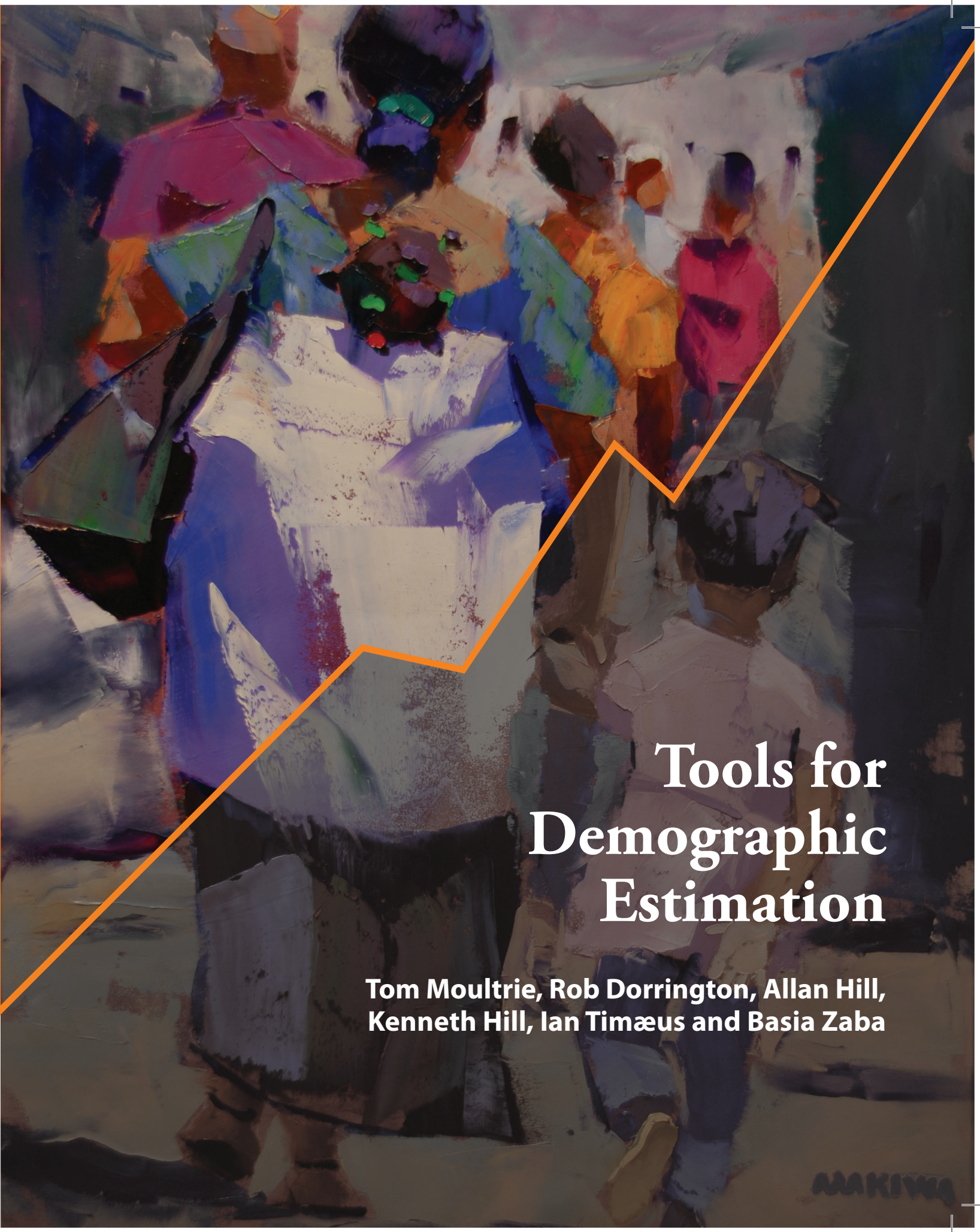


# Tools for Demographic Estimation

Tom Moultrie, Rob Dorrington, Allan Hill,
Kenneth Hill, Ian Timæus and Basia Zaba

Tools for Demographic Estimation

# Tools for Demographic Estimation

Tom Moultrie, Rob Dorrington, Allan Hill,
Kenneth Hill, Ian Timæus and Basia Zaba

3-5 Rue Nicolas, 75980 Paris Cedex 20, France
www.iussp.org



The suggested citation for this work is:
Moultrie TA, RE Dorrington, AG Hill, K Hill, IM Timæus and B Zaba (eds). 2013. *Tools for Demographic Estimation*. Paris: International Union for the Scientific Study of Population. demographicestimation.iussp.org

First published 2013
Second impression 2013

ISBN: 978-0-620-57491-4

Cover image: 'The Journey'. 2012. Makiwa Mutomba. Oil on canvas. 56 x 71 cm. Reproduced by kind permission of the artist.

Design and typesetting: User Friendly, Cape Town
Printed and bound by Paarl Media Paarl, 15 Jan van Riebeeck Drive, Paarl, South Africa

# Contents

# Introduction

*Tools for Demographic Estimation* is the result of a project, funded by the United Nations Population Fund (UNFPA) and run under the auspices of the International Union for the Scientific Study of Population (IUSSP), to bring together in one place, and in a user-friendly style, key methods used by demographers everywhere to measure demographic parameters from limited and defective data.

The idea for *Tools for Demographic Estimation* first arose at a joint IUSSP/UNFPA meeting on 'Applied and Technical Demographic Training in Developing Countries' held in The Hague in March 2009, where concern was expressed that the training of demographers in the use and application of indirect estimation techniques was waning at almost every academic institution around the globe.

Several factors have contributed to this state of affairs. First, changing global population priorities, notably the revised agenda adopted by the International Conference on Population and Development held in Cairo in 1994, had altered the funding landscape, with more resources being devoted to the emergent fields of reproductive and sexual health rather than the technical demography required to study patterns of growth and to manage population increase. Associated with this, the cohort of demographers who had been trained in the classical methods and techniques was ageing rapidly and few younger demographers were being trained in either the science or the craft of demographic estimation from limited and defective data.

Second, the Demographic and Health Surveys programme (DHS), associated with the collection of full birth histories and attendant direct estimation methods for fertility and mortality, has created the impression that the tools and techniques for estimating mortality and fertility from census or other survey data were no longer as important as they had been in the past. While there can be no doubt that the DHS has contributed enormously to, and helped reshape, the discipline of demography, the growing marginalization of demographic analysis of census data and other demographic materials limits our ability to understand demographic dynamics in developing countries. The role of the census in providing a sampling frame for demographic surveys is often forgotten. Moreover, the typical sample size of most DHS means that precise estimates from such surveys are seldom available at spatial resolutions smaller than regions or provinces, while the information collected on relatively rare events (such as adult deaths) is usually too sparse to permit the derivation of robust estimates.

Third, in most parts of the developing world (sub-Saharan Africa being the notable exception), improvements in systems of vital registration and the collection of demographic data in censuses mean that the existing techniques of demographic estimation from limited and defective data are regarded as obsolete. It is certainly the case that in countries with complete and accurate registration of vital events and a series of reliable censuses, direct and continuous estimation of demographic parameters becomes possible. In many low-income and middle-income countries, however, neither condition yet prevails and so it remains important to evaluate critically the quality of registration-based statistics and cross-check them against census-based questions on fertility and mortality.

A further reason for the decline in the priority accorded to the teaching of indirect techniques of demographic estimation is the natural evolution of populations where even in the poorest countries, fertility is falling after several decades of mortality improvement. The age distributions of these populations are thus far from the theoretical stable or even quasi-stable population model so that many of the techniques based on such models and first formally published by the UN Population Division (1967) are clearly outmoded. This demise of so-called stable population analysis led some analysts to prefer the DHS-style direct estimation methods over the whole suite of methods developed initially by Ansley J Coale and William Brass, authors of the early United Nations volumes.

In many instances, direct demographic estimation from census, survey, or vital registration data remains impossible

or problematic. This implies a continuing need for census-based and other indirect estimates. The 2009 meeting further noted that the canonical manual for demographic estimation from census data, *Manual X* (UN Population Division 1983), was more than a quarter of a century old and that several new methods and techniques had been developed since its publication. Two other manuals have been prepared since, *Estimating Demographic Parameters from Census Data* (Sloggett, Brass, Eldridge et al. 1994) and *Methods for Estimating Adult Mortality* (UN Population Division 2002) but neither attempted a full and comprehensive revision and update of *Manual X*.

The meeting at The Hague therefore resolved that a project be initiated to revise and update *Manual X*. Following a competitive call for proposals evaluated by the IUSSP, a consortium of demographers based at the University of Cape Town and the London School of Hygiene & Tropical Medicine and independent demographers associated with Harvard University was awarded the contract to develop the material. *Tools for Demographic Estimation* is the result.

The material presented here follows in a direct line of descent from *Manual X* and the rationale underpinning the work is fundamentally the same – to set out the methods for estimating demographic parameters from limited or defective data. We therefore strongly urge users of *Tools for Demographic Estimation* to read the introductory chapter to *Manual X* (available on the UN Population Division website) both for its description of the need for and history of indirect estimation methods, and for its discussion of the limitations of reference works of this kind.

*Tools for Demographic Estimation* differs from its precursors in several important respects. The differences stem in part from the enormous increases in computing power available to analysts since the time *Manual X* was published. They also reflect advances in approaches to demographic estimation, new methods, and the evolution of insights into how well different methods work, and under what conditions. Thus, the methods described in *Tools for Demographic Estimation* and the earlier manuals are not the same. A number of methods that have been developed since the publication of *Manual X* are presented here for the first time. Other methods that were presented in *Manual X* have been excluded on the grounds that they have since been found to work poorly or that more refined or newer methods render them obsolete.

Second, unlike its precursors, *Tools for Demographic Estimation* is primarily an electronic, web-based, resource. The print version represents the material on the project's website (*demographicestimation.iussp.org*) at the date of printing. The website, however, is designed to be dynamic, updated and changing over time. It follows that, whenever possible, the reader's primary point of reference should be the website, rather than the print version of the manual. The website, hosted by the IUSSP, is freely and readily accessible to anyone on registration.

Third, the website includes downloadable spreadsheets that implement the methods described, so as to facilitate their application and use. The decision to implement the methods using spreadsheets rather than in the form of downloadable executable programmes (such as, for example, *MortPak*) is intended to ensure a maximum degree of transparency. The formulae and calculations are visible to the end-user, and the spreadsheets can be modified by users if they do not exactly match the data available. The spreadsheets are in Microsoft Excel format but have been designed to be compatible with other open-source spreadsheet applications. Only in exceptional circumstances have Excel-specific facilities (such as *Solver*) been employed.

A fourth difference from earlier manuals on indirect estimation is that while *Tools for Demographic Estimation* adopts much the same approach as its precursors in providing step-by-step descriptions on how to apply the methods covered, a greater degree of emphasis has been placed on setting out the assumptions underlying each of the methods, as well as the situations and conditions under which the methods may be contra-indicated, or may produce unreliable results. To assist users interested in understanding how the methods work, we have endeavoured to present the mathematical derivation of the methods in as accessible a style as possible.

Fifth, *Tools for Demographic Estimation* incorporates material on the assessment and measurement of migration using census data, an area not covered at all in *Manual X*, and last described in a work of this kind in *Manual VI* (UN Population Division 1970).

Despite these advances, the present work suffers from many of the same limitations as its precursors. In presenting each method separately, the bigger picture associated with demographic estimation from limited and defective data is all too often lost. A significant component of this kind of demographic work lies in piecing together a puzzle composed of demographic parameters from multiple

methods and sources into a coherent, internally-consistent whole. Demographic estimation of the kind presented here is, ultimately, as much a craft as it is a science. Where possible, we have sought to give a sense of the craft involved. To facilitate and encourage the careful application of the methods described here, the website also includes discussion forums, which we hope will provide a vehicle for discussion of the results from applications of the methods presented, for suggestions for modifications or corrections to existing methods, and for proposals for new approaches to demographic estimation from limited and defective data.

*Tools for Demographic Estimation* has been a work long in preparation. The editors record their gratitude to the many people and organisations that have helped bring the project to fruition. We note the contributions of Ralph Hakkert (UNFPA) and Mary Ellen Zuppan (IUSSP) in securing funding for and overseeing the project; of the anonymous reviewers appointed by the IUSSP who offered extensive and useful comments on the initial draft of the material; of the web designer (Charles Oertel) and book designer (Jo-Anne Friedlander); and of the proof-reader (Debbie Budlender). We are also exceedingly grateful to those responsible for the UN Manuals as well as the Statistical Institute for Asia and the Pacific for waiving copyright and allowing us to reproduce material from those resources where necessary.

**Tom Moultrie, Rob Dorrington, Allan Hill, Kenneth Hill, Ian Timæus and Basia Zaba**
Cape Town, July 2013

## REFERENCES

Sloggett A, W Brass, SM Eldridge, IM Timæus, P Ward and B Zaba (eds). 1994. *Estimation of Demographic Parameters from Census Data*. Tokyo: Statistical Institute for Asia and the Pacific.

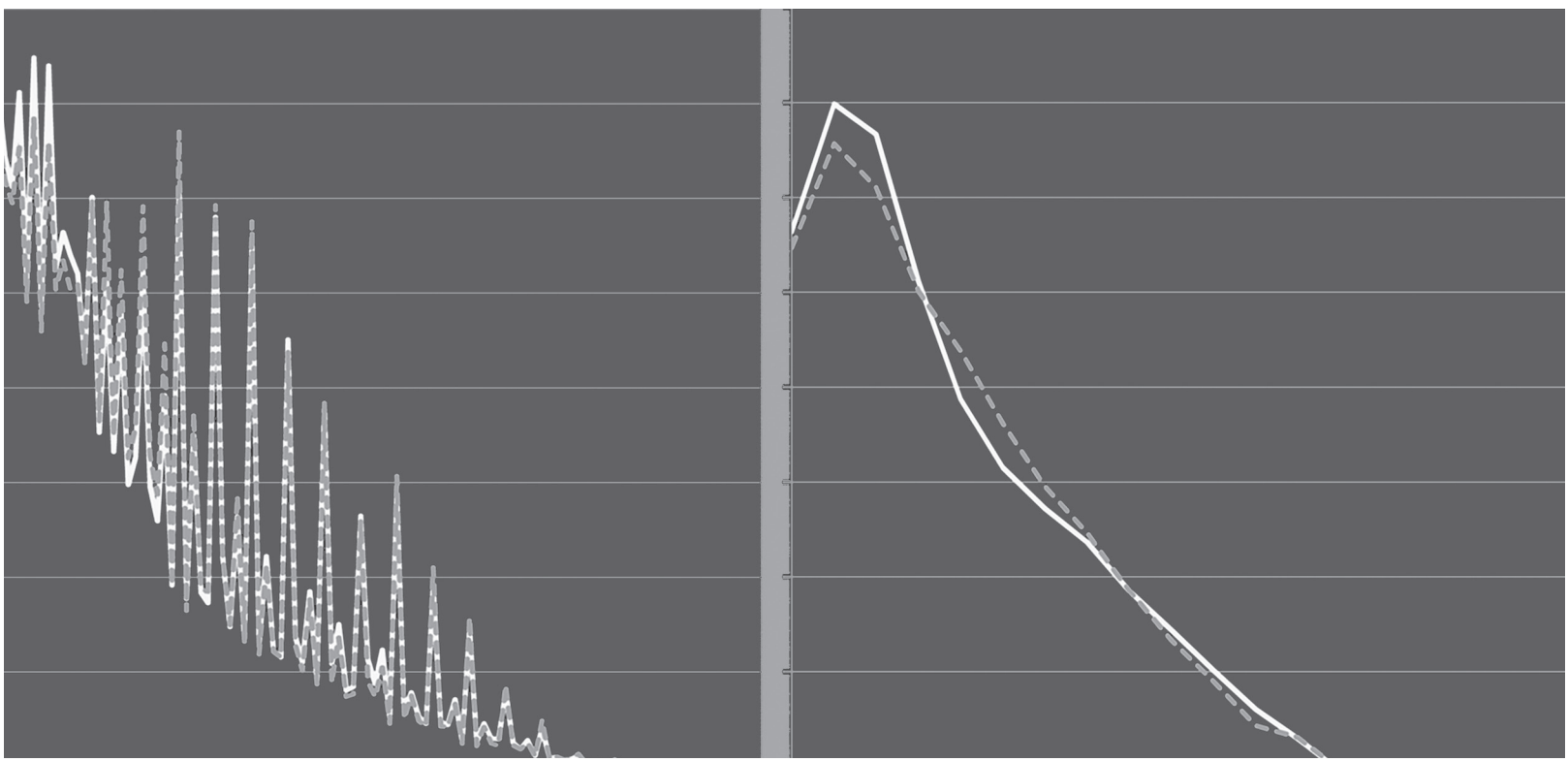
UN Population Division. 1967. *Manual IV: Methods for Estimating Basic Demographic Measures from Incomplete Data*. New York: United Nations, Department of Economic and Social Affairs, ST/SOA/Series A/42. http://www.un.org/esa/population/techcoop/DemEst/manual4/manual4.html

UN Population Division. 1970. *Manual VI: Methods of Measuring Internal Migration*. New York: United Nations, Department of Economic and Social Affairs, ST/SOA/Series A/47. http://www.un.org/esa/population/techcoop/IntMig/manual6/manual6.html

UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

UN Population Division. 2002. *Methods for Estimating Adult Mortality.* New York: United Nations, Department of Economic and Social Affairs, ESA/P/WP.175. http://www.un.org/esa/population/techcoop/DemEst/methods_adultmort/methods_adultmort.html

# Evaluation and Assessment of Data



# Chapter 1 General assessment of age and sex data

*Tom A Moultrie*

## INTRODUCTION

"In a perfect world, data would always be complete, accurate, current, pertinent, and unambiguous. In the real world, data are generally flawed on some or all of these dimensions" (Feeney 2003: 190). The task of evaluating and assessing data is an essential part of identifying the nature, direction, magnitude and likely significance of these flaws. While the primary point at which data evaluation and assessment takes place is immediately after the data have been processed, data evaluation and assessment are recursive activities – at each analytical stage, the user of demographic data should consider the results produced with a sceptical eye, alert to possible indications of error or bias introduced by the data into the results.

Here we set out the essential investigations that should be carried out as a matter of course before embarking on a process of demographic analysis. The basic principles for performing demographic evaluation and assessment have barely changed in the last half-century. Accordingly, aspects of the material presented in this chapter have been drawn from the United Nations' *Manual II: Methods of Appraisal of Quality of Basic Data for Population Estimates* (UN Population Branch 1955), updated and modified as appropriate, as well as from another, more recent, guide to the evaluation of census data, the United States Census Bureau's guide, *Evaluating Censuses of Population and Housing* (US Bureau of the Census 1985). The latter work provides a comprehensive and useful guide to the subject; in particular, Chapters 4 and 5 are strongly recommended to all analysts setting out on a process of data assessment and evaluation of demographic data.

The next section explains why it is necessary to evaluate demographic statistics. It also provides a high-level overview of the principles and practices involved.

## THE NEED FOR APPRAISAL OF DEMOGRAPHIC STATISTICS

Population statistics, like all other demographic statistics, whether they are obtained by enumeration, registration, or other means, are subject to error. The errors may be large or small, depending on the obstacles to accurate recording which are present in the area concerned, the methods used in compiling the data, and the relative efficiency with which the methods are applied. The importance of the errors, given their magnitude, depends on the uses to which the data are put. Some applications are valid even if the statistics are subject to fairly large errors; other applications require more accurate data. When dealing with any given problem, it is important to know whether the data are accurate enough to provide an acceptably accurate answer.

For population estimates, evaluation of the census or registration statistics on which the estimates are based is doubly important. In the first place, an investigation of the accuracy of the base data is a prerequisite to any attempt at determining the reliability of the estimates. Errors of estimation result both from inaccuracies in the basic population statistics and from errors in the assumptions involved in deriving the estimates (for example, in the assumed population changes between the date of the latest statistics and the date to which the estimate applies). Both sources of error must be taken into account if the degree of confidence that may be placed in the estimate is to be known. Second, where an investigation into the accuracy of the base data has revealed errors, the direction and magnitude of which can be estimated, it is possible to make explicit or implicit compensating adjustments, as the estimates of population are prepared. It is often the case, too, that reasonably reliable demographic measures (for example, fertility rates) can be derived, even when the underlying data are unreliable on some dimensions.

The purpose here is to describe the basic methods for

appraising the accuracy of those aspects of the census data most commonly used as a basis for current population estimates and future population projections. It is assumed that the results have been compiled from at least one census, and that the analyst is faced with the problem of determining the accuracy of the census and other population data, but is not in a position to re-enumerate the whole population or repeat any major part of the census undertaking.

It is not possible to consider in every detail all the possible information which, in a given country, can be utilized for an appraisal of its demographic data. For example, survey data may provide estimates of demographic parameters that would be valuable in evaluating the quality of data from a census. Hence, the examples presented here should be regarded as illustrations of methods and the results should not be taken as definitive evaluations of the quality of the particular data employed.

The results of the tests described in this manual are of various kinds. Sometimes, a test will reveal only that statistics are either "probably reasonably accurate" or "suspect"; if they are "suspect" further intensive investigation is required before a definite judgment can be made. Other tests will not only indicate that errors are present, but also lead to an estimate of the direction and probable extent of the error. In the latter case, it is desirable to adjust or correct the faulty statistics and to revise the estimates based on them. The description of procedures to be used in the revision of estimates, however, is outside the scope of this manual.

The distinction that is often drawn in demographic texts between coverage errors (introduced through differential enumeration across regions, ethnic groups, ages etc. leading to the data set being unrepresentative of the statistical whole it is meant to represent) and content errors (introduced through respondent or enumerator error, or misreporting) is not particularly helpful in determining strategies for data assessment. In many instances flaws in the data may not be attributable solely to one or the other kind of error. However, in seeking to explain and understand the errors identified, it is useful to consider where in the census process the error may have been introduced. Doing so assists in the determination of appropriate remedial courses of action to correct the data if possible. The description of such remedies is again outside the scope of this manual.

## BACKGROUND DOCUMENTATION THAT SHOULD BE SOUGHT

The process of conducting a census is arduous and complicated – it has been claimed, for example, that the decennial census conducted in the United States is the largest and most complex peacetime undertaking of the Federal Government (National Research Council 2004). The same is probably true in any other country conducting a census. To assist with the task, recommended standards and procedures have been drafted by the United Nations Statistics Division. Many of the relevant manuals are available online: the two of greatest interest to demographers analysing and evaluating the quality of census data are the *Principles and Recommendations for Population and Housing Censuses, Revision 2* (UN Statistics Division 2008) and the *Handbook on Population and Housing Census Editing, Revision 1* (UN Statistics Division 2010a). The former offers guidance on the logistics of conducting a census, from planning all the way through to dissemination; the latter deals with the post-enumeration handling of the data in preparation for release.

The nature and quality of the demographic data available varies greatly between countries. Population censuses are undertaken with varying frequency and accuracy, and vital registration data contain widely divergent levels of detail, and vary hugely in quality between and within countries. Migration across national boundaries may be relatively important or not. Consequently, different methods have to be employed in different situations for the appraisal of the accuracy of statistics, and it is therefore not possible to consider all the detailed tests to which every conceivable kind of data on the subjects covered here can be submitted. The methods presented here may, therefore, not always be directly applicable to a specific problem; modifications must be identified to suit particular requirements.

Where possible the analyst should seek to obtain as much relevant information as possible from the agency responsible for conducting the census or survey regarding operational practices and difficulties experienced, as well as the policies and practices adopted for cleaning and editing the data prior to release. Where a post-enumeration study has been conducted, information on this should also be obtained.

In addition to data sources that may not be in the public domain, the quality of the insights gained into the nature of the data will depend on the ability of the analyst to bring to bear on the data as much potentially relevant material, not only demographic, but also social, economic, historical and

political information as possible. As a simple example (more on which below), the dramatic decline in adult survival probabilities in the late 1990s indicated by the data from the 1992 and 2002 Zimbabwean Censuses can be explained in large measure by the effects of HIV/AIDS on adult mortality at that time.

## TYPES OF TESTING PROCEDURES

Whether one is dealing with census data, vital statistics, or records of migration, the same basic types of testing procedures are applicable. This similarity arises from the fact that demographic phenomena are interrelated both among themselves and with other social and economic phenomena. Some of these relationships are direct and necessary. For example, the increase in population during a given interval is precisely determined by the numbers of births, deaths, and net migratory movements occurring in that interval. Other relationships are less precise and less definite. For example, in some countries, an economic depression is likely to result in a declining, and prosperity in a rising, birth rate, but the exact amount by which the birth rate will change cannot be inferred even from detailed knowledge of the economic situation.

The basic types of possible testing procedures can be summarized as follows:

a) consistency checks, based on one or more censuses;
b) comparison of observed data with a theoretically expected configuration, for example the use of balancing equations and population projection models;
c) comparison of data observed in one country with those observed elsewhere;
d) comparison with similar data obtained for non-demographic purposes; and
e) direct checks (re-enumeration of samples of the population etc.).

The first type of checking procedure examines the consistency of the data, either internally (for example, does the distribution of the population by age and/or sex conform to expectations), or externally by means of comparison with earlier data from the same country. Demographic transition theory leads us to expect that – typically – birth rates and death rates (and hence population growth rates) will decline in a coherent, orderly fashion, without major discontinuities. (The exception is the likelihood that, at the very start of the transition, birth rates may rise). In the absence of clearly identifiable exogenous factors (e.g. war, famine or epidemics), deviations and departures from this orderliness therefore strongly suggest problems in the data.

Comparisons of the second type have changed significantly over the years. Historically, the most common tests of this type were to compare the data against those implied by a stable-population equivalent of the country in question. With the onset of fertility decline in almost every country in the world, the assumptions necessary for comparisons of this type to produce meaningful results have become increasingly invalid. Contemporary comparisons of this sort now more frequently seek to compare male and female mortality rates and sex ratios by age with those that would be anticipated in contexts similar to those of the source of the data being investigated. In addition, comparison with the results of model outputs (for example, the United Nations' World Population Prospects or the US Census Bureau's projections) can be used to highlight possible inconsistencies in the data.

"Balancing equations" can also be applied to test the consistency of the increase in population shown by two enumerations at different dates, using the increase shown by statistics of the various elements of population change – births, deaths, and migration – during the interval. If all the data were accurate, the two measures of increase (or decrease) should be balanced. Aside from population totals, the test can also be applied to sex and age groups and other categories of population that are identifiable in the statistics. Furthermore, by rearranging and re-defining the components of this equation, separate appraisals can be made regarding the accuracy of birth, death and migration statistics.

The third type of test relies on prior knowledge of a country that is expected to be demographically similar to the country of interest. This may, for example, be a neighbouring country. However, great care must be taken if this approach is to be adopted to ensure that the similarities between the two countries are sufficiently great (not only demographically, but also socially, economically, culturally etc.) to permit the extrapolation of data from one demographic setting to another.

The second and third types of check are similar. The demographic changes observed in another country where conditions are presumed to be similar can sometimes be substituted for a theoretically expected configuration. In both cases, the comparisons will differ, whether by a large or a small amount. The essence of the test then rests on the answer to the question: Can the difference between

the observed and expected values be explained by historical events or current conditions in the country, the data of which are being tested? If not, then it must be concluded that the observed data are "suspect". Further investigation may yield an explanation of the difference, or it may furnish clear indications that the "suspect" data are indeed in error. Very often this kind of method is applied as a preliminary step, to suggest along what lines further testing should be undertaken.

The fourth type of test relies on the availability of administrative or other social statistics that may shed light on the demography of the country of interest. Estimates of the sizes of different components of a national population might be obtained from voters' registers, school enrolment statistics, select populations such as Demographic Surveillance Sites (DSSs), etc. If such estimates differ from the population census data, the question arises whether there is a satisfactory explanation for the difference. Given the dependency on the specifics of the local data available, and the nature of the comparisons that might be drawn, tests of this type are not discussed further here. However, care must be taken not to assume these alternative sources are necessarily better than the census being checked.

Finally, direct checks involve a field investigation, such as a post-enumeration survey. The advantage of a direct check consists in the fact that the individual persons enumerated, or the individual events registered, can be identified, so that not only the consistency of totals, but the specific errors of omission or double-counting come to light. Direct checks in the form of a post-enumeration survey also allow for the correction of the enumerated population for an estimated undercount.

The first four types of testing procedure give an indication only of relative accuracy as both sets of data may be subject to error. If several testing procedures are applied, or if there is a strong presumption that one set of data used in the comparison is highly accurate, the evidence so secured provides a strong indication that the data being tested are inaccurate. In many other instances, the comparison may only reveal that at least one, if not both, sets of data are in error.

The investigations described below concentrate on the first and second types of test. (Direct checks are discussed briefly elsewhere, in the section on post-estimation consistency checks). Wherever possible, specific examples are included. The data for these examples have been drawn from the census data held at IPUMS (Minnesota Population Center 2010). However, only a fraction of the data and knowledge available in each country was used in working out these examples. Many more relevant data, some of them not published anywhere, exist in these countries.

A final observation before proceeding to the description of the various tests described here: most (although not all) of the tests can be applied at smaller geographical subdivisions, with the caveat that migration plays an increasingly significant role in determining the size and shape of populations at smaller levels of disaggregation. Here, too, we expect to find "orderly" patterns of population change, both within the same subdivision in successive intercensal periods, and among different subdivisions in any period. Any dissimilarities should be explicable in terms of known conditions. As a practical matter it is well known that there may be considerable diversity in the rates of population change among the various parts of any nation. Accordingly, the problem becomes one of trying to distinguish between changes which are explainable in terms other than errors in the statistics and those which are not.

## PRELIMINARY CHECKS

Before trying to assess the quality of the data the analyst should:

- Review the census enumeration procedures and information on the quality of performance, including ascertaining whether a post-enumeration survey was done, and whether the data should be weighted and, if so, how. Where possible, access to unedited, or only lightly-edited data should be sought, along with the manuals and algorithms used to edit the data.
- Ascertain how the data were collated into machine-readable form. Manual entry has the limitations of being slow; optical scanning – a technique adopted for many censuses in the 2000 and subsequent rounds of censuses – offers a faster processing time than manual capture, but is subject to numerous other faults (for example, difficulties in distinguishing 1s and 7s in many scripts), as well as problems associated with scanning the last pages of census forms, which may have become contaminated with dirt.
- Compare the census figures with any available data from non-demographic sources which relate to the numbers of the population or parts thereof.
- Compare the population distribution as revealed by the census findings to known characteristics of the

subdivisions; for example, the population density of rural areas should be less than that of urban areas.

- Compare the head and household counts (along with the average number of people per household and number of single-person households) at a national and regional level, and by urban/rural subdivisions to see if they make sense.

The degree of accuracy in a count of the total number of people in a country is directly related to the accuracy with which the entire census operation is conducted. The head count may be either more or less accurate than the enumeration of constituents of the population, such as by age or marital-status groups, but if all the census procedures are of poor quality and the characteristics of the population have not been accurately determined there is little likelihood that the head count will be correct. Indeed, one of the ways of appraising the quality of the head count consists of analysing the accuracy of data on various characteristics of the population. This analysis may not only reveal evidence of inaccurate classification of the individuals enumerated, but also may reveal a tendency to omit certain categories of the population. Special efforts should be made to appraise the completeness of the census counts in those areas or among those population groups which are known to be subject to conditions unfavourable for census taking. For example, there has been a long tradition of omission of very young children in censuses conducted across sub-Saharan Africa.

A detailed description of the factors which contribute to the completeness of a census count is beyond the scope of this manual. These factors are comprehensively discussed in many standard demographic texts (e.g. Shryock and Siegel (1976); UN Population Branch (1955)).

**Missing and edited data**

It is improbable that each and every respondent answered questions on both age and sex. If there are no missing data for these variables, the data have almost certainly been edited. Not all editing is bad. However, since a crucial part of determining the overall reliability of a data set hinges on the internal coherence of the age-sex structure of the population, it is preferable to be able to determine which data variables have been cleaned or edited as well as to be able to evaluate the rules applied to effect such changes. Sometimes this is indicated through inclusion of edit-flag variables, which may also indicate the types of editing and imputation that have been used for that particular variable. If this is the case, the distributions of the edited data according to the method used to derive the final data can highlight flaws or anomalies in the edit rules. Where possible, access to the unedited and uncleaned (or only very lightly edited/cleaned) data is desirable. Unfortunately, few countries release data with edit flags let alone provide access to a version of the data before editing took place.

The proportion of the data on any given variable that has been subjected to editing or imputation is also important. If too great a proportion of the data has been 'put' there by means of editing or imputation, the resulting distribution will reflect the assumptions underlying the rules used to edit the data rather than, necessarily, reality.

Where data on age are missing for some of the population, a decision needs to be made as to how to treat these records. Simply removing them from the analysis is not recommended: doing so reduces the absolute size of the population, and assumes that the age distribution of those people whose ages are missing is the same as that of those whose ages are not. If this is believed to be the case, missing ages in tables should be apportioned in accordance with the age distribution of the population whose ages are known. Thus (and analysing the data separately by sex, if required), if we define $N_x$ to be the enumerated population aged $x$, and $N_m$ to be the enumerated population with missing age, we would apportion these cases to individual ages:

$$N_x^* = N_x + N_m \frac{N_x}{\sum_{x=0}^{\omega} N_x} = N_x \frac{\sum_{x=0}^{\omega} N_x + N_m}{\sum_{x=0}^{\omega} N_x}.$$

However, if strong grounds exist to believe that the missing ages are clustered in a portion of the population, the apportionment should be modified to take this into account. For example, it may often be reasonable to assume that respondents would know the ages of children below a certain age, say 20.

When confronted with the need to apportion data on two dimensions (e.g. age and region), the approach set out by Arriaga in US Census Bureau (1997) should be followed. The method requires iteratively scaling the columns and then the rows to sum to the desired marginal totals. Convergence typically happens after a few iterations. The accompanying spreadsheet (see website) implements this approach and can handle up to 20 rows and 30 columns.

## CHECKS BASED ON THE AVAILABILITY OF DATA FROM A SINGLE CENSUS

The checks based on only one census should be done as a matter of course for all censuses, regardless of the availability of data from earlier censuses or surveys. These checks provide the basic insights into the demographic data collected in the census, and rest, largely, on evaluating the consistency and orderliness of the data by age and sex.

### Age- and sex- distributions

Given the centrality of age and sex in determining all three components of demographic change, investigations of the distributions of the population by age and sex are fundamental to any process of data assessment and evaluation. Investigations of this type can provide essential information on:

- the age and sex structure of the population;
- differential coverage or omission;
- the accuracy of reported ages, as well as the presence of digit preference; and
- whether the data have been subjected to editing or not.

#### *Population pyramids and other graphical assessments*

The drawing of population pyramids is not recommended as a tool for assessing the quality of demographic data, although they are useful for a number of other applications, and animated population pyramids are a useful instructional tool for demonstrating how populations change over time; cf. the examples of Canada or Germany). Historically, population pyramids were used to get a sense of the overall population structure as enumerated in the census. Although the graphing of rudimentary population pyramids in Excel is relatively straightforward, the correct formatting of them is laborious. More significantly, visual assessment of the data is difficult when the age-sex data are presented in this form. The same information (and more) can be far more readily provided simply by graphing the enumerated population by age and sex on the same pair of axes instead. The first assessment of the data should be done by single years of age, after which one can progress to examination of the five-year age distributions.

#### *Identification of heaping on age*

One of the benefits of graphing the population by single years of age and sex is that occurrences of data heaping by age are made visible from the start. Visual assessment of age heaping is probably as good an indication of age heaping as those of derived measures such as Myers' Blended Index, Whipple's Index or the United Nations Age-Sex Accuracy Index. These indices can be useful for comparative purposes but the scales of the indices are indicative at best, and the added information gained from the index over a simple graphical assessment often does not justify their use. The US Census Bureau's manual reaches a similar conclusion: "While these procedures are useful as summary measures or for comparative purposes, they generally do not provide any insight into patterns of error in the data that cannot be obtained through graphical and ratio analyses of the data." (US Bureau of the Census 1985: 140)

Heaping usually – but not always – takes the form of concentrations of the age distribution of the population on ages ending in 0 or 5. Depending on how the age variable in the census is collected or derived, heaping may occur on other ages, too. For example, if age at the census is derived from the respondent's reported month and year of birth, heaping may occur on reported years of birth ending in 0 or 5 (1920; 1925, etc.); the associated heaping by years of age in completed years will depend on the census date. In addition, other forms of heaping may not be readily apparent – for example that occasioned by mass registration at one point in time, or events of major historical significance – leading to preferences for ages ending on 0 or 5 on that date.

Given the expectation of orderly demographic change in the absence of significant exogenous events, a smooth progression in the numbers of people enumerated at each age is expected. In developing countries where fertility has remained high, one would expect the population size to decrease monotonically by age. If the absolute number of births has been declining in recent years, one would expect to find fewer children at younger ages than at slightly older ages.

One limitation of graphing of the population by age and sex is that distortions and error in the data at older ages will be obscured by the (much) larger population sizes at younger ages. Ratios or relative rates can be used to explore possible distortions and errors for older ages. If no comparator data are available, then the higher age ranges should be considered separately.

#### *Age ratios*

While heaping on particular ages are generally more easily identified graphically than through calculated measures, the

calculation of age ratios can provide a useful indication of possible undercounts or displacements between age groups. The age ratio for a given age group is the ratio of twice the population in that age group to the sum of the population in each of the adjacent age groups. Algebraically,

$$ {}_nAR_x = \frac{2.{}_nN_x}{({}_nN_{x-5} + {}_nN_{x+5})}.100. $$

On the presumption that population change is roughly linear between age groups, the ratio should be fairly close to 100. Deviations from 100, in the absence of plausible exogenous factors (e.g. migration; past calamities affecting particular age groups) are indicative of undercount or displacement errors in the data.

An aberration in the population numbers in any one particular age group (either real, or arising from an error in the data) is likely to cause disturbances in the age ratios for the age groups on either side. If one age group is particularly small, this will result in the age ratio for that age group being below one, with spikes in the adjacent groups.

***Sex ratios***

A second class of checks is to assess the sex ratios in the population, both generally and at each age. The overall sex ratio (SR) is the ratio of the number of males per 100 females in the population. This ratio can then be disaggregated by age as follows:

$$ {}_nSR_x = \frac{{}_nN_x^m}{{}_nN_x^f}.100, $$

where ${}_nN_x^i$ represents the enumerated population of sex $i$ ($i = m$ or $f$) between ages $x$ and $x+n$.

Since female mortality is typically lower than male mortality in most populations, the sex ratio should reflect this mortality differential. In developed countries, the sex ratio at birth (SRAB, the number of male children born per 100 female children) is typically around 105, while in sub-Saharan Africa, it appears to be closer to 100 (Garenne 2004). Values of the SRAB, derived (for example) from the sex of the last reported birth in the census or vital registration data, outside this range are indicative of sex-selective abortion, infanticide, or reporting problems.

In the absence of significant net migration, the overall ratio reflects the relative mortality of females and males. Provided there are no specific reasons why female mortality might be higher than that of males (e.g. sex-specific foetal selection; infanticide of female babies; very high maternal mortality; or widespread neglect of women as discussed by Sen (1992)) one would expect the overall sex ratio to be slightly less than 100. Given the differences between male and female mortality, particularly at older ages, the exact magnitude of the overall ratio will be strongly conditioned by the age structure of the population, being lower for older populations, and higher for younger populations.

Between birth and late middle-age (around 45 in developing countries; 60 or older in developed countries) the sex ratio typically should decline only slowly unless there is significant net migration. Thereafter, the sex ratio tends to fall rapidly as male mortality begins to greatly exceed female mortality. A common departure from this pattern is visible in countries with high levels of sex-selective labour migration among young adults. If large numbers of young men are living outside the country at the time of enumeration, this will reveal itself in a sharp decrease in the sex ratios, followed by a gradual recovery among older men as these labour migrants return home.

**Concluding comments**

An integrated assessment of the quality of the data collected in a census and survey must seek to explain – with as few assumptions as possible – the features observed in the data. In this regard, the analyst must be alert to well-documented problems found with census data on age and sex – the undercount of young men of working age, and the exaggeration of ages that is frequently found in countries with some form of social welfare such as a state old-age pension. Finally, if there has been significant immigration, it may be useful to analyse the local-born population separately from the entire population; no comparable exploration is available for emigration, unless data by age, sex and country of birth are available for key destination countries.

**EXAMPLE**

The accompanying spreadsheet (see website) gives data from the 11.35 per cent sample from the 2001 Census of Nepal, held by IPUMS (Minnesota Population Center 2010). The data appear to have been subject to some kind of editing or cleaning, as there are no cases of missing age or sex in the data. The analyst should seek to determine the nature and extent of any such edits.

As suggested above, we begin by graphing the enumerated population by single year of age and sex (Figure 1.1).

Visual inspection of the left-hand panel immediately highlights the extreme digit preference for ages ending in 0 and 5 in these data. By way of example, the population of both males and females enumerated at age 30 is more than three times the population aged either 29 or 31. Heaping is also visible on ages ending in 2 and 8. Digit preference is less marked for the population aged less than 30, although this is in part due to the heaping visible on other ages (8, 12 and 18). Clearly the reporting of individual ages is not robust in these data.

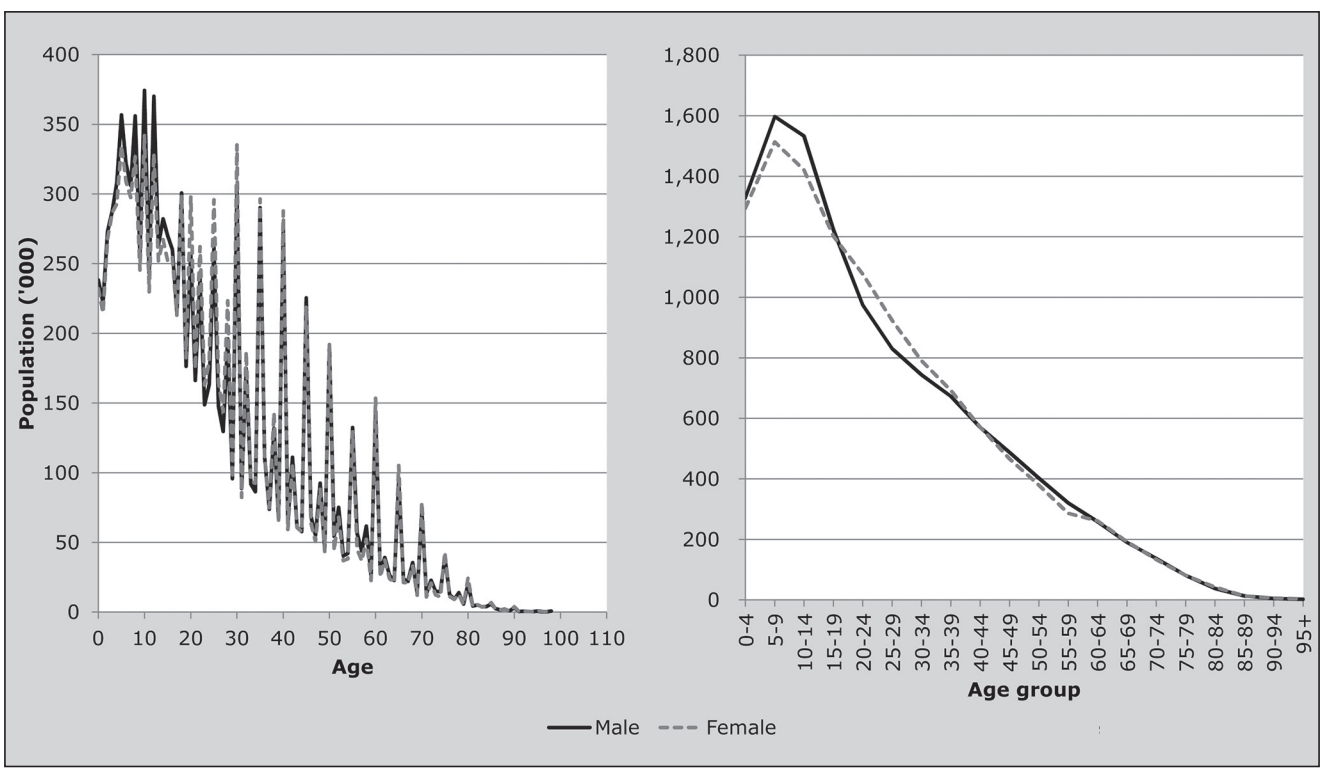
Also strongly evident in these data is the sharp fall-off in the enumerated population under the age of 5, with the enumerated population aged 1 being approximately two-thirds of that aged 5. It is unlikely that fertility has fallen by that magnitude in such a short period of time, and hence the initial presumption must be that young children were differentially undercounted in that census. Misreporting of children's ages – resulting in an over-statement of the number of children aged 5–9 might also have contributed to the shortfall of younger children.

Plotting the same age distribution in five-year age groups to smooth the data (right-hand panel of Figure 1.1) provides further insights. Again, the sharp fall-off in the population aged under 5 is visible, but visual comparison of the population aged 5–9 with that aged 10–14 suggests the possibility that there may have been some under-enumeration of children aged 5–9 too. This calls into question the possibility that there may have been large-scale transference of children aged 0–4 into the 5–9 age group. Finally, the age ratios for five-year age groups are shown in Figure 1.2.

The age ratios are generally close to 1 for both sexes, except at the youngest ages (indicating some omission of children aged 0–4, as well as a lesser degree of displacement of children into the 5–9 age group). The fall-off in the age ratios at the oldest ages is to be expected given the rapid increase in mortality at those ages.

In the absence of additional information, the age and sex distributions cannot be analysed further, but the analyst may wish to compare the relevantly aged population against administrative data indicating the numbers of children

*FIGURE 1.1 Age and sex structure, by single and grouped ages, Nepal, 2001 Census*



enrolled in school, or compare the administratively reported births 5–9 and 10–14 years prior to the census. A comparison can also be made against the estimated population aged 5–9 derived by applying estimates of fertility rates from the mid-1990s to the estimated female population at about that time.

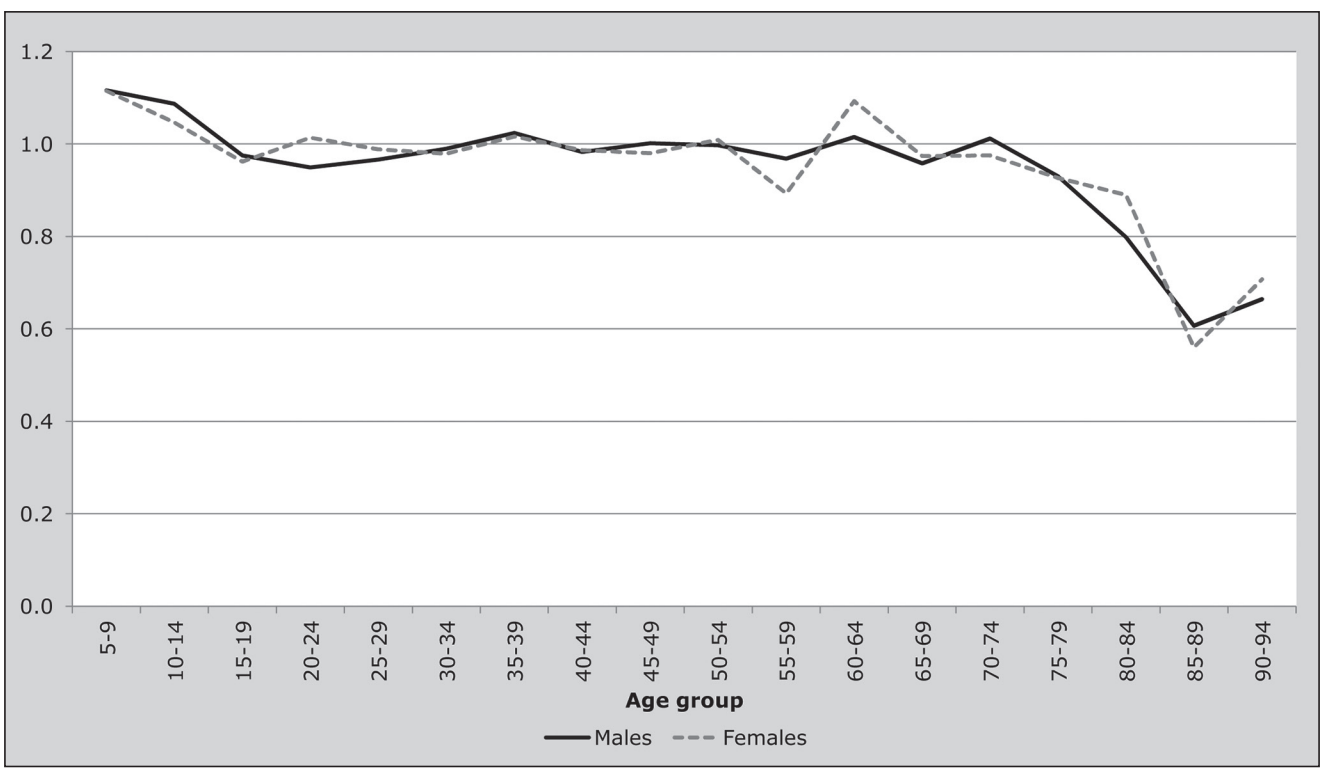
A second characteristic of the data that may require further investigation is the relative populations of males and females by age group. In aggregate the sex ratio of the enumerated census population is 100.5 men per 100 women. There is a noticeable surfeit of enumerated males until age 20. Between ages 20 and 40 there would appear to be more females than males. This could be the consequence of (male) labour out-migration, or a differential undercount of young adult men. The analyst should seek to find explanations for this phenomenon. However, (male) labour out-migration could plausibly account for some of the shortfall; the enumerated surfeit of men between the ages of 40 and 60 coincides with the ages at which men are most likely to return from work abroad, although this cannot account for the sex ratio rising above unity. One explanation might be that the sociological phenomena (sex-selective abortion; female infanticide) described by Sen (1992) in India might apply equally in Nepal.

Considering Figure 1.3, two features of the sex ratios by single years of age (left-hand panel) stand out. First, they are quite erratic, falling sharply from age 60 onwards at ages ending in 0 and 5. This suggests that ages of men were less likely to be heaped on those digits, and more likely to be heaped for women. Second, in addition to the deficit of men between the ages of 20 and 40 identified earlier, judging from the fact that the sex ratios remain above (or very close to) unity until the oldest ages, there would also appear to be a shortage of women over the age of 40 in the census. Again, the applicability of Sen's hypothesis to Nepal should be investigated.

The data presented by five-year age groups (right-hand panel of Figure 1.3) is smoother, but nonetheless reaffirms the analysis above.

Further insights into the nature and quality of the age and sex data from the 2001 Nepal Census can be gained from a

***FIGURE 1.2*** *Age ratios by sex and five-year age groups, Nepal, 2001 Census*



comparison of these data with the United Nations Population Division's most recent estimates for the country in 2001 (UN Population Division 2011). These estimates stand in marked contrast to the census data. The most effective way to show the differences is to plot the ratio of the enumerated population (by age and sex) to the UN Population Division's estimated population for 2001 (Figure 1.4).

Ratios above the age of 90 are not shown as they are even more extreme – rising to 9.7 (for males) and 8.6 (for females). If they were shown, they would mask the differences at younger ages.

While the UN estimates primarily reflect the assumptions that went into them, the huge discrepancies between the two sets of estimates require careful investigation. Up until age 15, the ratios for males and females follow almost identical trends. However, the enumerated population of males and females at age 0–4 is some 30 per cent lower than that estimated by the UN, while that at 10–14 is within two or three per cent. At older ages, the patterns by sex diverge markedly: the number of women between the ages of 15 and 55 differs between the two data sources by an almost-constant five per cent.

Relative to the UN projections, there appears to be extensive age exaggeration at older ages, especially amongst men.

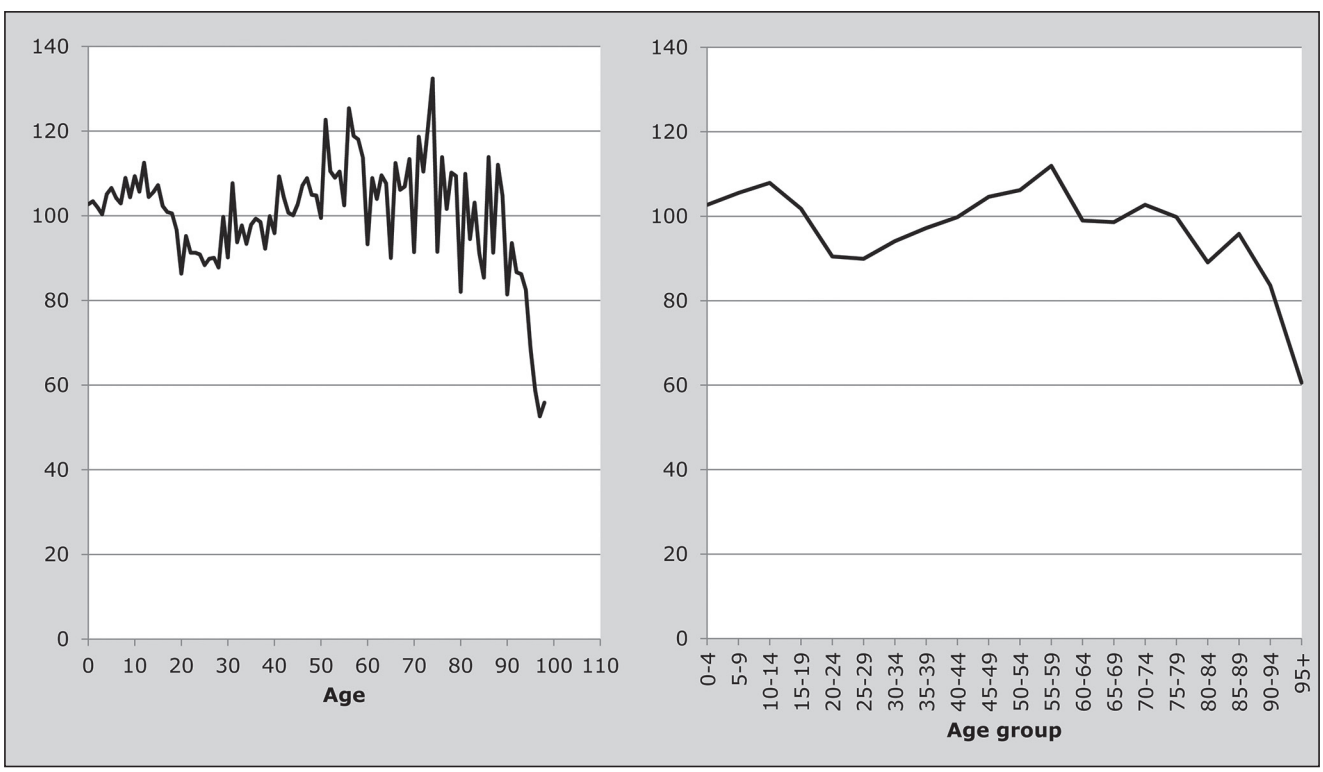
The comparison of the sex ratios by age calculated from the 2001 Census data, and those estimated for 2001 by the UN Population Division (Figure 1.5) also reveals noticeable differences. Further work is certainly required to understand what may account for the widely divergent accounts of the demographic structure in this country.

## CHECKS BASED ON MULTIPLE CENSUSES

In addition to the checks described in the previous section, the availability of additional sets of data from earlier censuses (and vital registration systems) makes other investigations possible.

It is often difficult to determine whether irregularities revealed by the evaluation of the age and sex structure of a population in a single census are due mainly to errors in the data or to real peculiarities of the population structure.

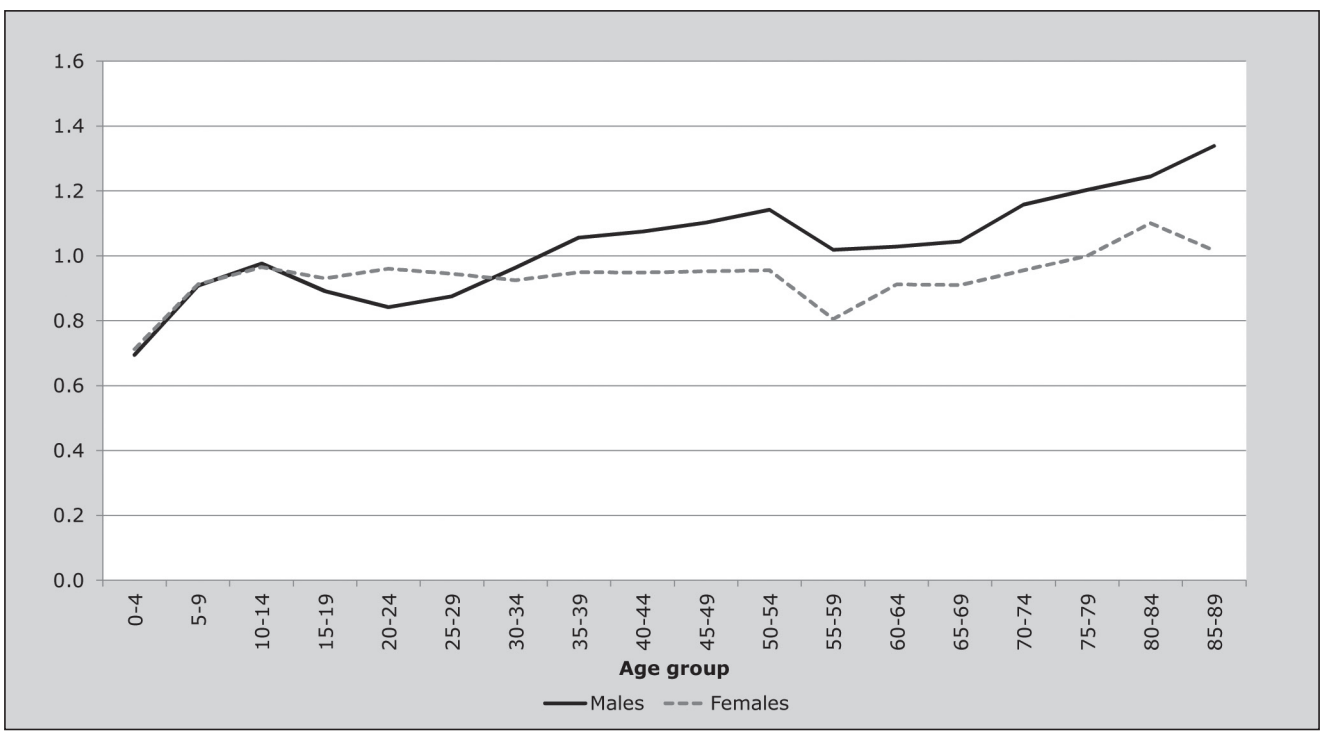
*FIGURE 1.3 Sex ratios by single and grouped ages, Nepal, 2001 Census*



When the results of two or more successive censuses are available, it is often possible to clear up these uncertainties even without the use of any more elaborate techniques than were described in the preceding section. For example, if the age statistics from the 2008 Census of Cambodia were at hand, the possibility of explaining certain irregularities in the 1998 data as the results of birth deficits or deaths in the period of the Khmer Rouge's rule in the late 1970s would be greatly clarified. If the 2008 figures should show the same peculiarities in the age groups ten years higher, but not in the same age groups in which they appeared in 1998, there would be a strong basis for concluding that these peculiarities reflected the true figures, rather than enumeration errors. Still more definite information regarding errors can be obtained where data from two or more censuses at intervals of a few years are available, by using balancing equations or analogous calculations with the data for particular cohorts – comparing, for example, the numbers reported at ages 10–14 in an earlier census with those reported at ages 20–24 in a census ten years later. Where data from a series of three or more censuses are available, the returns may be linked in this manner over the entire series. For the purpose of explaining the techniques, however, it is sufficient to consider examples of the use of data from two censuses.

Again, the guiding principle to be followed in comparing the results from two or more successive censuses is that population changes normally proceed in an orderly manner. When such an orderly pattern is not observed, the deviations should be explainable in terms of known events, such as the curtailment of immigration, the occurrence of famine, or some other event. Deviations from the pattern which cannot be so explained constitute a warning of possible errors; and the presumption of error is greatly strengthened if the results of other tests are found to point in the same direction. In some countries it may be possible to apply these tests to the various ethnic groups separately, if age and sex data are tabulated for such groups and if data are available on immigration and emigration of these groups (or if the groups in question are not substantially affected by international migration).

***FIGURE 1.4*** *Ratio by age and sex of enumerated population in 2001 to UN Population Division World Population Prospects (2011) estimates for 2001, Nepal*



Checks making specific use of multiple censuses are, for the most part, based on (and in some cases, are) methods used to measure adult mortality – in other words, the assessment of the consistency of the data is a by-product of the methods to estimate adult mortality. This chapter describes some of these based on the data that are likely to be available.

**Evaluation of intercensal growth rates**

The growth rate, $r$, is defined as

$$r = \frac{\ln\left(\frac{N(t_2)}{N(t_1)}\right)}{(t_2 - t_1)},$$

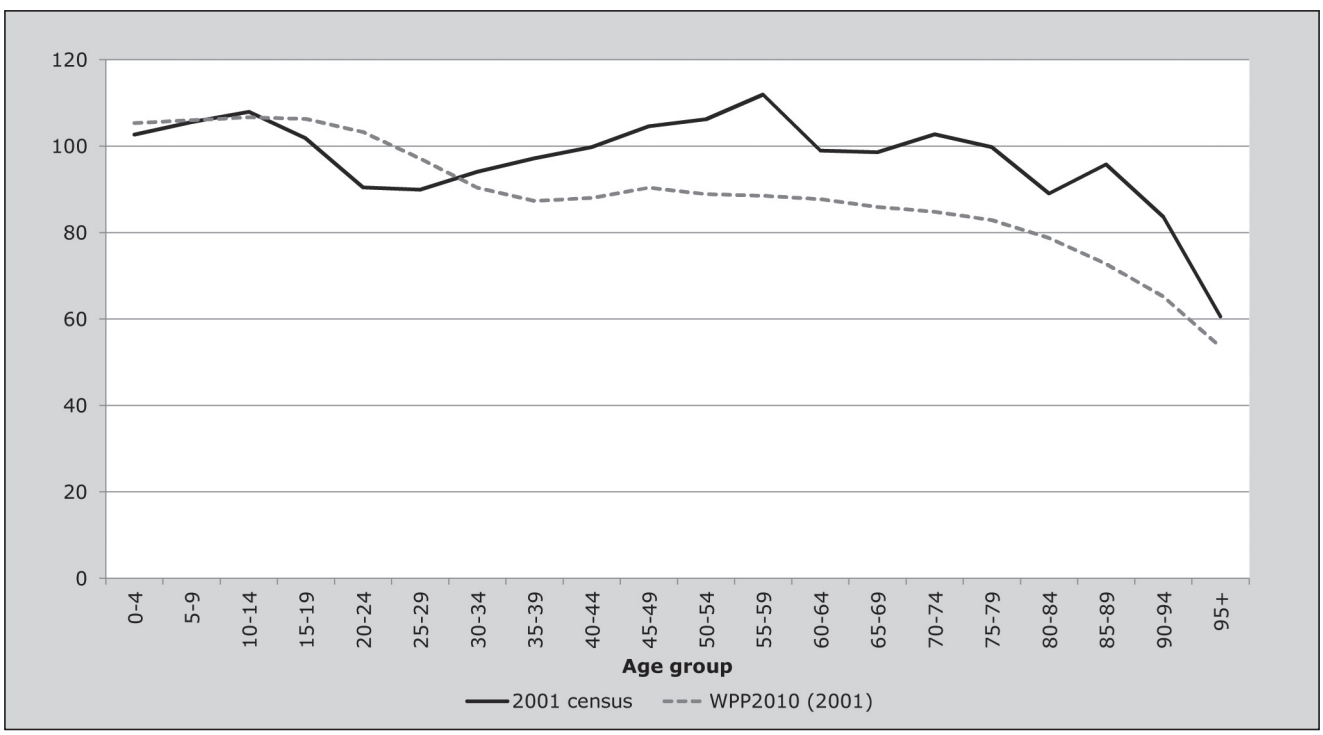
where $N(t_1)$ is the total population at time $t_1$, and similarly for $N(t_2)$.

If the country's population changes only through natural increase, it is very unlikely to have an average annual rate of growth exceeding 3.5 per cent. A rate at this level would be the result of a high birth rate (say 45 or more per 1,000) and a very low death rate (say 10 or less per 1,000). Further, it is only in unusual circumstances that the population of a developing country would be likely to decline without heavy emigration. In fact, nearly all observed rates of natural increase in modern times have been in the range from zero to 3.5 per cent. In a few developed countries, according to the 2010 World Population Prospects, natural growth is negative. If, in any given country, the rate of population change approaches or exceeds these limits without large-scale immigration or emigration, the question must be raised as to whether there is some explanation for such an unusual rate, or whether the census counts were in error.

With some information, however approximate, regarding the conditions of mortality and fertility in the country, the limits of the likely rate of growth may be defined more closely.

If population counts are available for three or more successive censuses it becomes possible to make a more accurate evaluation by comparing successive rates of growth. Again, the same principle is followed, namely, that the

*FIGURE 1.5 Sex ratios by age group from the enumerated population in 2001 and the UN Population Division World Population Prospects (2011) estimates for 2001, Nepal*



pattern of population growth should be regular except in so far as it can be shown that changes in the circumstances may have led to departures from the pattern.

Further, provided the censuses are undercounted to the same extent, the estimate of *r* is correct. Thus tracking *r* can provide an indication of relative undercount between pairs of data.

**Cohort survival ratios**

Any particular age group can be defined as a cohort: for example, boys under 5 years of age, women 50 to 54 years, or all persons 10 to 19 years of age, at a given census date. If a second census is taken exactly one decade later, the surviving members of each cohort will be exactly ten years older at the time of the second census. However, their numbers will be reduced by deaths and they may be increased or reduced by the balance of immigration and emigration. Ordinarily, mortality is the main factor; if the migration balance is negligible, the change in numbers can be used to compute a survival ratio analogous to that of a life table. Computed for one cohort only, such a survival ratio often reveals little, if anything, about the accuracy of the statistics. However, a patently absurd result would give clear evidence of error. For example, an increase in the numbers of a cohort, from one census to another, is obviously impossible, unless there has been a substantial amount of immigration. Similarly, even under conditions of very high mortality, it is unlikely that a cohort aged anywhere between 5 and 60 years at the beginning, will be reduced by one-half within a decade.

More accurate judgement is possible if the survival ratios are compared for cohorts of each sex at different ages. Survival ratios are functions of age-specific death rates, and, like these, generally conform to more or less the same pattern of variation from age to age whether mortality is high or low. The rate of survival increases after the earliest years of childhood and usually attains its maximum around age 10; thereafter it declines, at first very gradually, but more and more rapidly as advanced ages are attained. Also, at most or all ages females usually have a somewhat higher rate of survival than do males of the same age. If the hypothetical survival ratios computed for different cohorts deviate significantly from this pattern, and if no explanation (such as migration) can be found, inaccuracy in the statistics must be suspected.

Under what conditions can such comparisons of cohorts in successive censuses be made most meaningfully? One condition is either the absence of substantial net immigration or emigration or full knowledge about the age and sex composition of the migrants. A second condition, analogous to the first, is that of constant boundaries. If the country's boundaries have changed between the two censuses so that considerable numbers of people have been added to or subtracted from the population, the age and sex composition of these people must be known, if the cohort analysis is to give an accurate indication of the accuracy of the statistics. A third condition is that the population covered by the two censuses must be the same. For example, if the entire male population is enumerated in one census, but the military is excluded at the second, the age cohorts involving the military cannot be compared without a suitable adjustment, unless the number of the military is negligible. If nationals living abroad are included in one census, and excluded from another, and if the numbers involved are large, especially if they are concentrated in any particular age or sex groups, this type of analysis is invalidated.

In the case of a country where immigration is substantial, under certain circumstances a cohort can be compared at two censuses even if migration data are lacking. If the native-born population (that is to say, persons born in the country) are known not to have emigrated in significant numbers, comparisons of the two censuses can be limited to that population.

Survival ratios can be calculated over any age span and time interval, provided one has data by single years of age for at least one of the pair of censuses. With the decennial programme of censuses recommended by the United Nations, a ten-year span of ages is typical.

***Method***

Cohort survival ratios (CSR) measure the proportion of people enumerated at age $x$ to $x+n$ at time $t$, ${}_nN_x(t)$, in the first census, who are still alive and enumerated in a second census $a$ years later when they are aged $x+a$ to $x+n+a$ at time $t+a$, ${}_nN_{x+a}(t+a)$. Thus

$$ {}_nCSR_x(a) = \frac{{}_nN_{x+a}(t+a)}{{}_nN_x(t)}. $$

For graphical presentation, these estimates can be located at the mid-point of the intersurvey period (i.e. at time $t+\frac{n}{2}$) and at the midpoint of the ages at that time, $x+\frac{a+n}{2}$.

A plot of these cohort survival ratios offers easy and rapid insights into the quality of the data at hand, although the standard caveat still applies; a curious sequence of cohort survival ratios indicates that something may have gone amiss with the data, but does not indicate whether the fault lies with the first, the second or both censuses.

Where data from a third census are available, however, it may be that the cohort survival rates derived from the first two censuses appear reasonable, while those derived from the second and third appear problematic. In this case, one would proceed by assuming that the fault lies with the enumeration in the third census, and not the second.

Finally, if one has an appropriate life table at one's disposal, one could derive a further ratio by dividing the Cohort Survival Ratio into the equivalent ratio implied by the life table, resulting in a ratio of ratios at each age

$$ {}_nR_x = \frac{{}_nN_{x+a}(t+a) \big/ {}_nN_x(t)}{{}_nL_{x+a} \big/ {}_nL_x}. $$

If the census suffered no error, the age structure of the enumerated population was identical to that described by the life table, and the mortality experience was exactly that indicated by the life table (all three strong conditions), the ratio would take the value of 1. Departures from unity would indicate either error in the data, or an inappropriate choice of life table. Further, under these conditions and in the absence of migration, ratios less than unity would imply an under-enumeration in the second census relative to the first and vice versa.

**Example**

Censuses in Zimbabwe were conducted exactly 10 years apart (the official Census date being 18 August) in 1982, 1992 and 2002. Tabulations of the enumerated population by age and sex are available from the Demographic Yearbooks on the UN Statistics Division Website (the 1997 Historical Supplement and the 2008 Yearbook were used). The tabulations are shown in Table 1.1. The data present the population aged under 1 and aged 1–4 separately; these populations are kept distinct for the purpose of more accurately understanding the population dynamics given the rapid changes in mortality in the first few years of life.

Provided there are no grounds for believing that records with missing ages are not concentrated disproportionately in certain age groups, the first step is to apportion the (proportionately relatively small) number of cases where age

*TABLE 1.1 Population of Zimbabwe by age and sex, 1982, 1992 and 2000 Censuses*

| | 1982 | | 1992 | | 2002 | |
|---|---|---|---|---|---|---|
| Age | Male | Female | Male | Female | Male | Female |
| 0 | 133,070 | 136,960 | 167,552 | 169,064 | 170,054 | 170,277 |
| 1–4 | 510,260 | 528,390 | 621,411 | 626,664 | 668,008 | 667,730 |
| 5–9 | 612,760 | 619,300 | 821,319 | 832,469 | 764,453 | 769,247 |
| 10–14 | 529,750 | 518,740 | 724,905 | 731,846 | 754,587 | 757,657 |
| 15–19 | 390,160 | 412,610 | 615,728 | 632,510 | 736,686 | 766,890 |
| 20–24 | 290,380 | 364,200 | 466,837 | 523,060 | 564,034 | 658,873 |
| 25–29 | 243,420 | 281,060 | 335,713 | 376,495 | 473,984 | 513,793 |
| 30–34 | 185,400 | 206,760 | 280,066 | 326,299 | 369,836 | 360,291 |
| 35–39 | 147,920 | 170,170 | 229,360 | 259,555 | 235,692 | 268,797 |
| 40–44 | 142,050 | 139,530 | 174,266 | 189,509 | 194,702 | 239,727 |
| 45–49 | 116,490 | 110,390 | 145,437 | 143,441 | 165,437 | 191,168 |
| 50–54 | 111,780 | 90,880 | 133,261 | 147,339 | 128,029 | 173,229 |
| 55–59 | 67,400 | 60,800 | 94,713 | 86,729 | 98,417 | 112,498 |
| 60–64 | 76,850 | 65,260 | 95,510 | 84,213 | 94,447 | 99,420 |
| 65–69 | 38,810 | 38,860 | 51,202 | 50,902 | 64,301 | 67,851 |
| 70–74 | 29,810 | 30,500 | 58,279 | 62,479 | 60,311 | 62,464 |
| 75+ | 39,410 | 46,760 | 52,026 | 68,403 | 71,950 | 92,311 |
| Unknown | 7,900 | 6,680 | 15,952 | 18,034 | 19,252 | 25,254 |
| **TOTAL** | **3,673,620** | **3,827,850** | **5,083,537** | **5,329,011** | **5,634,180** | **5,997,477** |

is missing in proportion to the population size in each age group from 0 to 75+. In the 1982 Census, the proportion of the population with unknown age was 0.19 per cent; this doubled to 0.38 per cent in the 2002 Census. The resulting distributions are shown in Table 1.2.

In keeping with the principles outlined earlier, the basic age and sex characteristics of the population are investigated first. The unavailability of data by single years of age means that this aspect of the data quality cannot be investigated. The age and sex distributions of the Zimbabwean population from the three censuses are shown in Figure 1.6.

In all three censuses there is a clear surfeit of women between the ages of 15 and (at least) 35. This is almost certainly a product of labour migration of young men to neighbouring countries, most obviously South Africa. There would appear to have been a sizeable underenumeration of the population under the age of 5 in the 1992 Census – the population in that age group is less than that then aged 5–9, unlike the adjacent censuses.

Age and sex ratios from the three censuses are shown in Figure 1.7. The age ratios in the 60–64 age group are particularly high in all three censuses, and the excess population at that age contributes to the very low age ratios in the 55–59 and 65–69 age groups. The sex ratios start off close to 100, and fall rapidly after age 15 in each census, probably indicating migration of young men. Of greater concern is the rise in the sex ratios between ages 35 and 55 to levels far in excess of 100 in the 1982 Census. This almost certainly reflects some undercount of women. Sex ratios at the oldest ages are still very high, probably reflecting age exaggeration among older men.

The highly erratic age and sex ratios do not inspire a great deal of confidence in the quality of the data.

Next, cohort survival rates are derived as above, for each sex separately since patterns and level of mortality differ for males and females. Since the population aged 0–4 in 1992, for example, would be aged 10–14 in 2002, we assume that the survival rate for this cohort applies (roughly) to people aged 7½ at the midpoint between the censuses in August 1997. Cohort survival rates are not estimated for the very young, or for the open interval. The results are presented graphically in Figure 1.8.

The top left panel shows the cohort survival rates between the 1982 and 1992 Censuses by sex; the bottom left panel shows the equivalent data from the 1992 and 2002 Censuses. There was an evident undercount of children of both sexes as well as women up until around the age of 20 in the 1982 Census (or, improbably, high levels of child in-migration

*TABLE 1.2 Adjusted population of Zimbabwe by age and sex, 1982, 1992 and 2002 Censuses*

| | 1982 | | 1992 | | 2002 | |
|---|---|---|---|---|---|---|
| Age | Male | Female | Male | Female | Male | Female |
| 0 | 133,357 | 137,199 | 168,079 | 169,638 | 170,637 | 170,997 |
| 1–4 | 511,360 | 529,314 | 623,367 | 628,792 | 670,298 | 670,554 |
| 5–9 | 614,081 | 620,383 | 823,904 | 835,296 | 767,074 | 772,500 |
| 10–14 | 530,892 | 519,647 | 727,187 | 734,331 | 757,174 | 760,861 |
| 15–19 | 391,001 | 413,331 | 617,666 | 634,658 | 739,212 | 770,133 |
| 20–24 | 291,006 | 364,837 | 468,307 | 524,836 | 565,968 | 661,659 |
| 25–29 | 243,945 | 281,551 | 336,770 | 377,773 | 475,609 | 515,966 |
| 30–34 | 185,800 | 207,121 | 280,948 | 327,407 | 371,104 | 361,815 |
| 35–39 | 148,239 | 170,467 | 230,082 | 260,436 | 236,500 | 269,934 |
| 40–44 | 142,356 | 139,774 | 174,815 | 190,152 | 195,370 | 240,741 |
| 45–49 | 116,741 | 110,583 | 145,895 | 143,928 | 166,004 | 191,976 |
| 50–54 | 112,021 | 91,039 | 133,680 | 147,839 | 128,468 | 173,962 |
| 55–59 | 67,545 | 60,906 | 95,011 | 87,023 | 98,754 | 112,974 |
| 60–64 | 77,016 | 65,374 | 95,811 | 84,499 | 94,771 | 99,840 |
| 65–69 | 38,894 | 38,928 | 51,363 | 51,075 | 64,521 | 68,138 |
| 70–74 | 29,874 | 30,553 | 58,462 | 62,691 | 60,518 | 62,728 |
| 75+ | 39,495 | 46,842 | 52,190 | 68,635 | 72,197 | 92,701 |
| **TOTAL** | **3,673,620** | **3,827,850** | **5,083,537** | **5,329,011** | **5,634,180** | **5,997,477** |

between 1982 and 1992), as indicated by the survival ratios greater than one.

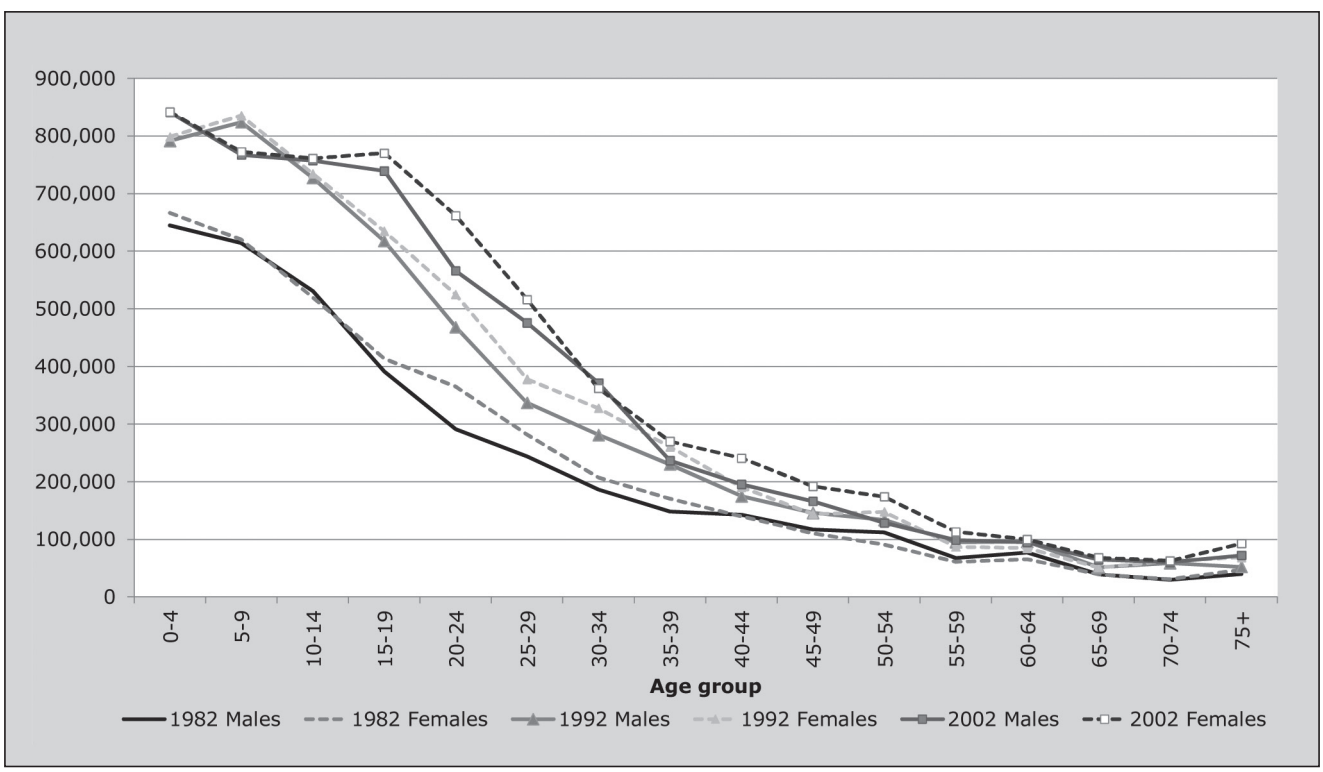
While both left-hand panels show (broadly) a pattern of decreasing survival ratios (increasing mortality) by age, the data are far from consistent either by sex or by age. It is unlikely, for example, that the survival ratios for men will be greater than those for women of the same age. There is also a very curious spike in both intercensal periods in the survival ratios of women aged 40–44 in first period to 50–54 in the next. This should be investigated further.

The two right-hand panels depict the cohort survival ratios over time, for men and women separately. They indicate a pattern of substantially increased mortality in Zimbabwe over the two ten-year periods. While the erratic nature of the survival ratios indicates the relatively poor quality of the data, this increase is almost certainly attributable largely to the effect of HIV/AIDS among adults in the country in the second period, in conjunction with the rapidly worsened socio-economic conditions that prevailed in the country towards the turn of the century, which almost certainly fuelled extensive out-migration of younger adults. The apparent increase in mortality among children and young adults seen in the two right-hand panels is almost certainly largely attributable to the poor enumeration of this population in 1982.

## POST-ENUMERATION SURVEYS

A post-enumeration survey (PES) uses the logic of capture-recapture techniques to estimate the proportion of the population that was not enumerated at the time of the census. This is done by returning to sample sites to readminister a second, short, questionnaire to all households which should have been enumerated in that site, after which households and individuals captured in this survey are matched, wherever possible, with those from the census. This procedure should give a concrete estimate of the magnitude of the undercount which can be compared to and contrasted with that implied by, for example, an analytical population projection. The results from the PES, then, can be used to scale up ("weight") the enumerated data to compensate for the undercount.

*FIGURE 1.6 Age and sex distribution of the Zimbabwean population, 1982, 1992 and 2002 Censuses*



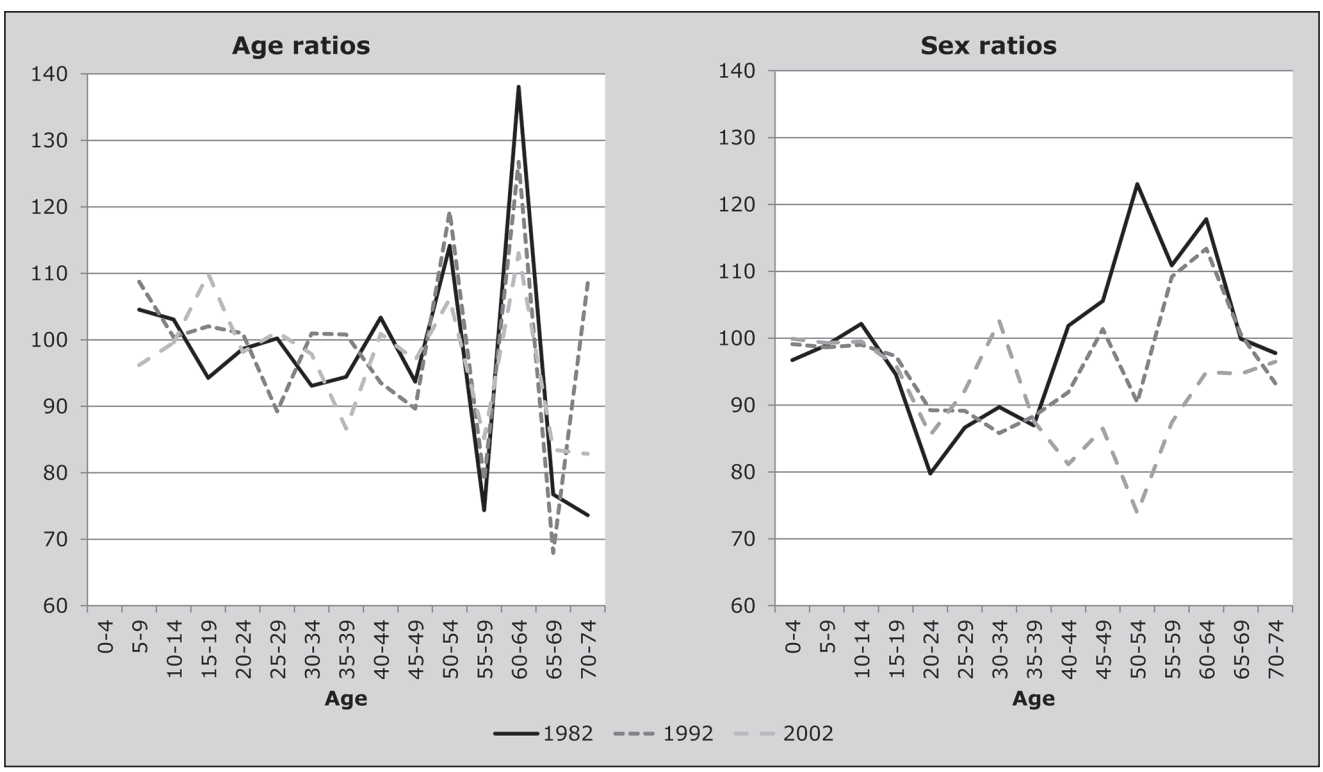
A PES is thus potentially extremely useful. However, the two key assumptions underlying the use of capture-recapture techniques are that the probabilities of being found in the census and the PES are independent of each other; and that it is possible to identify the same individual unambiguously in both data sets. The first assumption is unlikely to hold in human populations – certain groups who avoid being counted in a census (illegal immigrants, for example), are likely also to avoid a PES. In this sense, the PES gives information only on those known to have been missed in the census; it tells nothing of those not known to have been missed. The second assumption is also unlikely to hold, particularly in settings with high population mobility or if the interval between the census and the PES is long.

The principles and best practice associated with the conduct of a PES are documented in a 2010 manual (UN Statistics Division 2010b).

Where a PES is conducted, ideally the analyst will have access to the report on the PES so as to understand any deficiencies in that study. The ability of the PES to provide finely-grained insights into the data collected in the census is directly related to the size of the PES as well as to the delay between the census date and the date of the PES. Given the time and cost constraints, the sample size of a PES is typically much smaller than a full census. Accordingly estimates of an undercount have to be made at a fairly coarse level. Thus, for example, in the 2001 South African Census, the estimates of the undercount were made using only broad age groups, sex, population group, province, and enumerator area geo-type (urban, rural, formal, informal). In turn, this means that the population is assumed to be equally undercounted within each group defined by the five characteristics above. Hence, at granularities finer than those used to determine the undercount, the resulting estimate of the count may not be reliable.

Insights into the magnitude of the adjustments made, and the extent of the undercount, can be gained from an evaluation of the weights provided with the data. If the raw data made available from a census are unadjusted by a PES, then the data weights will reflect the sampling fraction: in a

***FIGURE 1.7*** *Age ratios and sex ratios, Zimbabwe 1982, 1992 and 2002 Censuses*



random 10 per cent sample drawn from a full census, each record would be deemed to be representative of 10 people, and hence carry a weight of 10. Where a PES has been conducted, the excess of the weight over the sampling fraction reveals the undercount. Analytically,

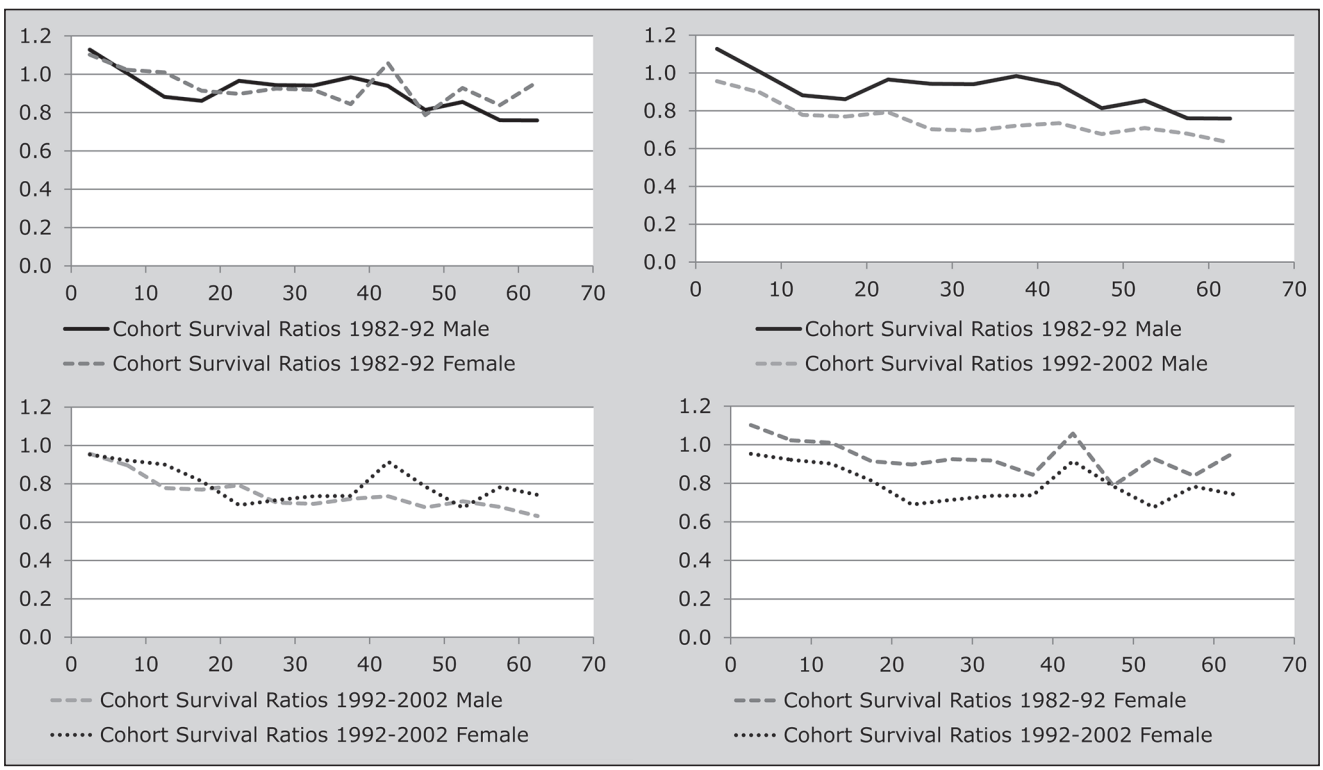
$$weight = \frac{sample\ fraction^{-1}}{1 - undercount}$$

$$undercount = 1 - \frac{sample\ fraction^{-1}}{weight}.$$

Hence, if for a particular record, the weight is 11.8 in a ten per cent sample (i.e. a sample fraction of 0.1), this implies an adjustment in respect of an undercount of 15.3 per cent (1–(1/0.1)/11.8).

Where estimates of the undercount have not been provided with the data, applying this last formula to the weights provided to different groups within the population allows the analyst to reverse-engineer the estimated undercounts to a fairly high degree of accuracy.

***FIGURE 1.8*** *Cohort survival ratios by age and sex, 1992 and 2002 Zimbabwean Censuses*



## REFERENCES


Feeney G. 2003. "Data assessment," in Demeny, P and G McNicoll (eds). *Encyclopaedia of Population*. Vol. 1. New York: Macmillan Reference USA, pp. 190–193.

Garenne M. 2004. "Sex ratios at birth in populations of Eastern and Southern Africa", *Southern African Journal of Demography* **9**(1):91–96. http://www.commerce.uct.ac.za/Organisations/Demography/SA_Journal_of_Demography/SAJD/Volume%209/SAJDem_2004_9_1_Garenne_b.pdf

Minnesota Population Center. 2010. Integrated Public Use Microdata Series, International. Version 6.0 [Machine-readable database]. Minneapolis: University of Minnesota.

National Research Council. 2004. *The 2000 Census: Counting under Adversity.* Panel to Review the 2000 Census. Citro, CF, DL Cork and JL Norwood (eds), Committee on National Statistics, Division of Behavioural and Social Sciences and Education. Washington DC: National Academies Press.

Sen A. 1992. "Missing women", *British Medical Journal* **304**(6827):587–588. doi: 10.1136/bmj.304.6827.587

Shryock HS and JS Siegel. 1976. *The Methods and Materials of Demography (Condensed Edition).* San Diego: Academic Press.

UN Population Branch. 1955. *Manual II: Methods of Appraisal of Quality of Basic Data for Population Estimates.* New York: United Nations, Department of Economic and Social Affairs, ST/SOA/Series A/23. http://www.un.org/esa/population/techcoop/DemEst/manual2/manual2.html

UN Population Division. 2011. *World Population Prospects: The 2010 Revision, Volume I: Comprehensive Tables*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/313. http://esa.un.org/unpd/wpp/Documentation/pdf/WPP2010_Volume-I_Comprehensive-Tables.pdf

UN Statistics Division. 2008. *Principles and Recommendations for Population and Housing Censuses v.2*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/STAT/SER.M/67/Rev.2. http://unstats.un.org/unsd/publication/SeriesM/Seriesm_67rev2e.pdf

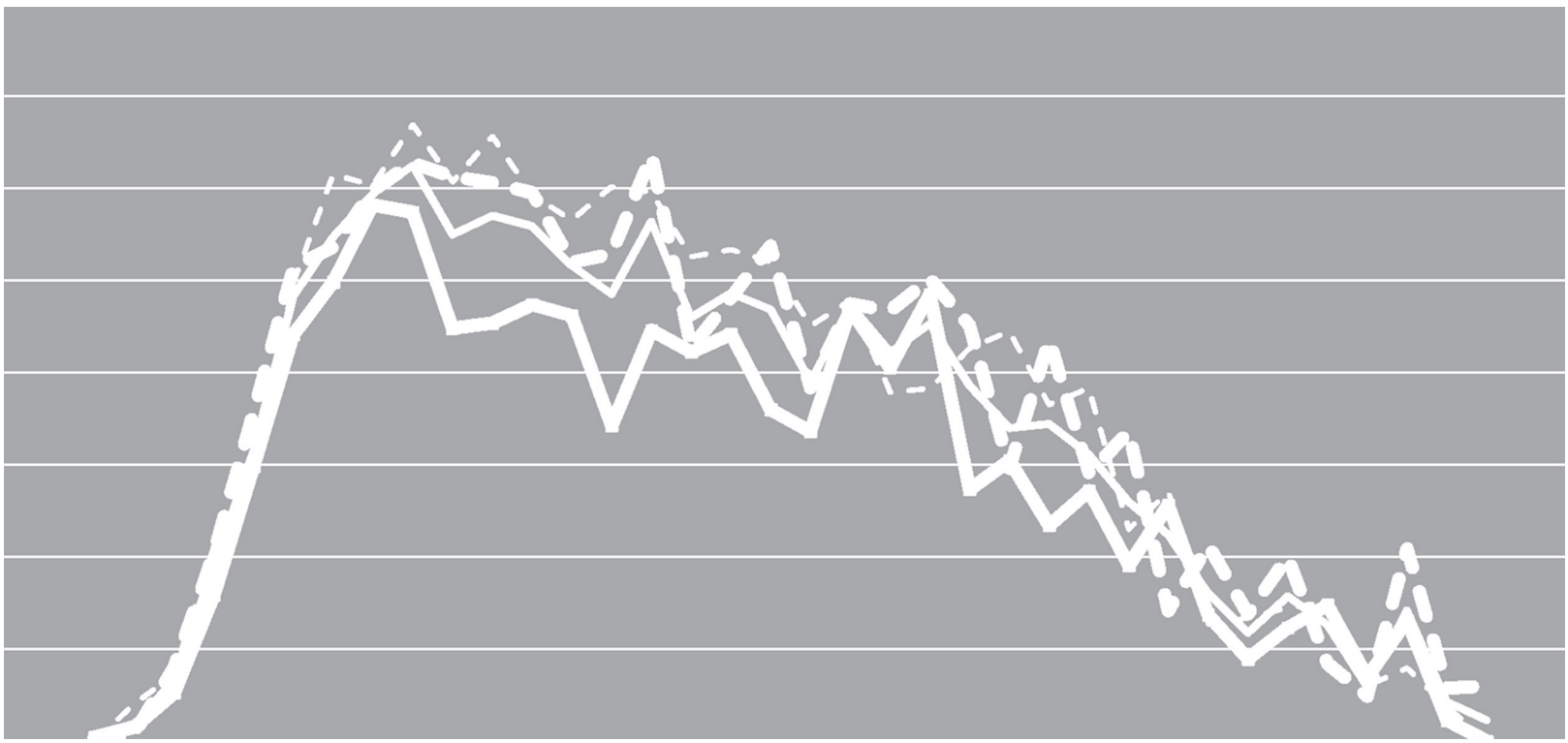
UN Statistics Division. 2010a. *Handbook on Population and Housing Census Editing, Revision 1*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/STAT/SER.F/82/Rev.1. http://unstats.un.org/unsd/publication/SeriesF/seriesf_82rev1e.pdf

UN Statistics Division. 2010b. *Post-Enumeration Surveys: Operational Guide lines – Technical Report*. New York: United Nations, Department of Economic and Social Affairs. http://unstats.un.org/unsd/demographic/standmeth/handbooks/Manual_PESen.pdf

US Bureau of the Census. 1985. *Evaluating Censuses of Population and Housing*. Statistical Training Document ISP-TR-5. Washington, DC: US Bureau of the Census. http://www.census.gov/srd/papers/pdf/rr85-24.pdf

US Census Bureau. 1997. *Population Analysis Spreadsheets for Excel*. Washington, DC: US Bureau of the Census. http://www.census.gov/population/international/software/pas/

# Fertility



# Chapter 2 Introduction to fertility analysis

*Tom A Moultrie*

In most settings and in the long term, fertility is the single most important determinant of population dynamics and growth. This chapter gives an overview of the classes of methods available for the measurement of fertility. The methods themselves are described in other chapters.

## SOURCES OF DATA

Broadly speaking, data for the measurement of fertility come from three sources.

The first of these is information collected as part of an ongoing vital registration system. For the most part, the registration of births in developing countries is incomplete: parents often lack incentives to register births; babies who die shortly after birth may not be registered either as a birth or as a death; and late registration of births (for example, when the child attains school-going age) may mean that there is a delay of several years before all the survivors of the cohort born in a given year have their births registered.

The second source of data is the answers to questions on fertility collected in a census. These questions are typically asked of all women of childbearing age (12, or 15 and older, often with an upper age limit of 49). Due to the complexity of the census exercise, and the need to attempt to enumerate every individual, it is not feasible to ask detailed questions on fertility. In practice, the census questions seek summary information about lifetime fertility (the number of children ever born, and still alive) and fertility in a narrowly defined period of time before the census. From the latter, demographers seek to estimate current fertility rates. Due to the abridged nature of the questions asked, the scope for internal validation and cross-checking of the answers given is limited.

Furthermore, as described in greater detail in the section on the evaluation of fertility data, data on fertility collected in censuses commonly suffer from two errors. First, data on lifetime fertility tend to be increasingly poorly reported with increasing age of the mother. Often the omissions are of children who have died or who are no longer living with the mother. Second, data on recent fertility tend to be systematically underreported by all women (similar to the widespread under-enumeration of the youngest children in the household head-count). Over-enumeration of recent births is also a possibility, occasioned by misunderstandings related to the reference period used, or shifting of the most recent births into the reference period. The methods used to estimate fertility from census data explicitly seek to take these errors into account.

Third, as well as asking these summary questions on fertility, surveys often collect a detailed birth history from mothers. Such histories ask about each child's date of birth, vital status (whether the child is still living) and – if the child is dead – date of death. The data obtained can be used to make detailed estimates of fertility. On the downside, however, the effort expended on such detailed data collection frequently limits the sample sizes of the investigations. In such cases, variability in the estimated rates, and the inability to investigate finely-grained spatial or other differentials in fertility are an inherent weakness of this approach.

## CLASSES OF METHODS TO ESTIMATE FERTILITY

The methods available to estimate fertility are closely aligned with the type of data available.

### Direct estimation

Three possible approaches to measuring fertility directly exist, depending on the data available.

The first approach uses data from a vital registration system in conjunction with estimates of the population by age and sex (from a population register, for example, or from mid-year population estimates). If data for the numerator and denominator are both complete and unbiased, and the denominator appropriately reflects the population exposed

to risk of giving birth by age, the calculation of fertility rates is straightforward.

The second approach makes use of the full birth history data collected in a survey and the various validation checks that can be built into the survey instrument. Detailed information on the birth of each child, as well as the mother's age, is obtained. Accordingly, the age of the mother at the birth of each child can be determined exactly, and births and exposure-to-risk can be allocated to particular calendar years or other time periods. This approach is described in greater detail in the chapter on direct estimation of fertility from survey data (Chapter 12).

The third approach uses the summary fertility measures routinely collected in censuses to estimate recent fertility. Methods for doing so are described in Chapter 5. Of course, if the data suffer from the common problem of under-reporting of recent births in censuses, the resulting estimates of fertility will be too low.

**Indirect estimation**

Indirect estimation of fertility makes use of the direct estimates of fertility derived from the summary information on recent births, but – recognizing that recent births tend to be misreported in censuses – uses information on the lifetime fertility of younger women reported in the same census to adjust the direct estimates. The earliest such method is the Brass *P/F* method, first set out by Brass (1964). The method was written up in *Manual X* (UN Population Division 1983), along with a number of variants for extending the method depending on exactly what data are available. A brief description of the *P/F* method is given in Chapter 6. A refinement to the *P/F* method – the relational Gompertz model – was mentioned in passing in *Manual X* but much of the development of the relational Gompertz model occurred after *Manual X* was published. This manual presents several extensions to the relational Gompertz model (the basic version of which is presented in Chapter 7) that emulate the extensions to the *P/F* method presented in *Manual X*. These include extensions to situations where:

- Lifetime and current fertility data from more than one census are available, and the analyst is seeking to estimate fertility for the intercensal period – the synthetic relational Gompertz model (Chapter 10).
- Only data on lifetime fertility are available – from two censuses or surveys conducted either five or 10 years apart – making it necessary to estimate fertility from the increments in parities. This is the parity increment method discussed in Chapter 11.
- Data on lifetime fertility are available from two censuses, along with information on births in the intercensal period from, for example, a vital registration system. This approach (presented in Chapter 14) allows one to assess the completeness of the data on registered births.

In this manual all methods of fertility estimation that make use of the *P/F* ratio method are recast here to use the relational Gompertz model.

**Further analysis of fertility**

Finally, there are several other methods that may shed light on fertility trends and dynamics.

Using census data, one can calculate both conventional and projected parity progression ratios. These measures indicate the propensity of women in a population to bear further children contingent on the number of children that they have already borne. Projected parity progression ratios indicate the possible future evolution of parity progression for younger women, taking into account current fertility and the women's childbearing history to date. These measures are presented in Chapter 8.

Second, methods exist (Chapter 13) for the calculation of cohort-period fertility rates from detailed birth history data. These rates not only provide information on trends in fertility, but can also be used to assess the quality of the birth history data.

Finally, there are methods for estimating fertility measures based on reverse survival of the enumerated population of children and adults (Chapter 9).

All the above-mentioned methods are described and discussed in this manual.

## REFERENCES AND FURTHER READING

The best earlier description of approaches to estimating fertility indirectly is contained in Chapter 2 of *Manual X* (UN Population Division 1983). The manual written by the Centre for Population Studies for the UN Statistical Institute for Asia and the Pacific (Sloggett, Brass, Eldridge *et al*. 1994) offers a simple exposition of evaluation of fertility data, the use of the relational Gompertz model in fertility estimation, and the calculation of projected parity progression ratios. The paper by Brass (1964) has been superseded by more recent developments, but is still instructive in its exposition of the approach which has shaped almost all subsequent methods for measuring fertility from limited and defective data.


Brass W. 1964. *Uses of census or survey data for the estimation of vital rates*. Paper prepared for the African Seminar on Vital Statistics, Addis Ababa 14–19 December 1964. Document No. E/CN.14/CAS.4/V57. New York: United Nations.

Sloggett A, W Brass, SM Eldridge, IM Timæus, P Ward and B Zaba (eds). 1994. *Estimation of Demographic Parameters from Census Data*. Tokyo: Statistical Institute for Asia and the Pacific.

UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

# Evaluation of Summary Fertility Data from Censuses

# Chapter 3 Assessment of parity data

*Tom A Moultrie*

The first type of question on fertility asked in censuses concerns women's lifetime fertility. It asks about their total number of live births. In order to reduce underreporting of dead or absent children (who are usually a larger proportion of children born to older women than younger women) and guard against underreporting of girls, the questions are often structured as a series of six questions about the number of sons and daughters:

- born alive and living with the mother;
- born alive but living elsewhere; and
- born alive but now deceased.

## TOTAL CHILDREN BORN AND SURVIVING

The total of the answers to the questions relating to living children, present and absent, provides the total number of children born and surviving. Adding the reported numbers of children dead gives the total number of children ever born to the woman. When summing these individual answers, care must be taken not to treat error or missing value codes as legitimate responses. For example, if a missing value is coded as '9', the procedure for deriving measures of the total children ever born, surviving and dead must make sure to exclude these codes.

Tabulations of the numbers of children reported in response to these questions are often truncated at some relatively high number (e.g. 9+). When this is the case, the only plausible assumption is that women in that category have had the number of children defined by the lower bound of the interval. The resulting errors are generally small, even in the case of extremely high fertility, unless the truncation is applied to the total children ever born, rather than to the separate categories of co-resident, absent, and dead sons and daughters.

### Implausible parities

In evaluating the quality of data on lifetime fertility, the analyst should be alert to improbable and implausible parities relative to the age of the mother. Especially at young ages, a small number of women reporting excessively high numbers of children ever born can have a material effect on the estimated mean children ever born. Such errors can result from misreporting, or manual or automatic mis-capturing of the data. A useful rule of thumb is to limit the maximum number of live births that a women may have had to one birth every 18 months from the age of 12, rounding down to the next integer. Using this rubric, by exact age 20 (the end point of the 15–19 age group), a woman might have had a maximum of 5 children; by exact age 25 (closing the 20–24 age group), 8. If the reported number of lifetime births exceeds this maximum, the recorded value should be recoded as 'missing'.

### Assessment of enumerator errors

Another common error in the recording of lifetime fertility is caused by the failure of the enumerator to record responses of 'zero' on the census form, leaving the relevant space blank instead. It is impossible to be sure whether a blank means that the enumerator omitted to ask the question or record the response or whether it indicates zero. This error is usually more common in the data on younger women, who are more likely to be childless or answer zero to some of the six questions above. The error in some cases occurs because the enumerator assumes that the question is not relevant for younger women, or feels uncomfortable about asking it. A specific adjustment to the data, the el-Badry correction, is often indicated in this case. However, if in every age group the number of women with unstated parity is low (as a guide, less than 2 per cent of the total), then this reporting error is unlikely to have a material impact on the derived average parities and these cases can be ignored in further calculations. This is the same as making the explicit assumption that

women with unstated parity have the same average parity as women in the same age group whose parity is known.

**Proportions of women childless**

The proportions of women who are childless should be calculated by age group of mother. The proportions should decline sharply with age. In most cases there should be around 3–10 per cent of women remaining childless in the oldest age-group, reflecting underlying levels of primary sterility and voluntary childlessness. In low fertility countries the proportion of childless women aged 45–49 may be even higher. Proportions of childless at older ages that exceed 10 per cent should be investigated further, as this may indicate significant errors in the data.

## AVERAGE PARITIES

A key indication of the consistency of data on women's lifetime fertility is a credible pattern of average numbers of children alive and dead by age group of mother. In general, one would expect average parities (the average total number of co-resident, absent and dead children born to women) to increase steadily with age. The shape of the distribution by age should be sigmoid, with slightly flatter sections at the beginning and end, reflecting lower fertility at the youngest and oldest ages at which women bear children. Significant parity increments in these age groups – that is, large increases in average parities between successive age groups – are unlikely.

One would also expect average numbers of living children, dead children and the proportion of children dead each to rise with age.

A second check is to compare the observed average parities with results from Demographic and Health Surveys (DHS), or from earlier censuses and other surveys. In this regard, one can compare the average parities for real birth cohorts of women. Thus if two censuses are conducted a decade apart, the average parities of women aged $x$ to $x+4$ in the earlier census can be compared with those of women aged $x+10$ to $x+14$ in the second. Average parities should not only increase monotonically with age within each census, but the cohorts should also show a reasonable parity increment between censuses.

If one has data on women aged 50 and over, one can make direct comparisons of the consistency of the average parities of women who have completed their childbearing – for example, by comparing the average parities of women aged 45–49 in one census with those of women aged 55–59 in a second census conducted a decade later. In making comparisons of this sort, and especially with comparisons involving older women, one should be alert to the possibility that mortality might differ according to the number of children a woman has had, either directly or because high fertility and socio-economic status may be correlated. This may hinder the ability to draw definitive conclusions about the trend in lifetime fertility.

A further refinement suggested by Feeney (1991) that is possible where there is information on the average parities of women who have completed their childbearing, is to locate these parities approximately in time and plot them. The approximate time location is derived by assuming that the average parities refer to a point time defined by subtracting the mid-point of each age group from the census date and assuming that all births in each cohort occurred at some mean age of childbearing, $m$. Thus, assuming $m = 27.5$ for example, if a census was conducted in 1960, the average parities of women aged 50–54 would refer (approximately) to $1960 - 52.5 + 27.5$, or 1935.

The average parity of women of a given age $x$, $P_x$, is calculated by dividing the total number of children ever born to women aged $x$ at the census date by the number of women aged $x$ at the census:

$$P_x = \frac{\sum_{j=0}^{\omega} j.N_{x,j}}{\sum_{j=0}^{\omega} N_{x,j}}$$

where $N_{x,j}$ is the number of women aged $x$ and of parity $j$ in the population, and omega ($\omega$) is the upper limit of the parities recorded in the population after excluding numerical values assigned as error codes in the data. In five-year age groups, the average parity of women in each age group is given by

$${}_5P_x = \frac{\sum_{j=0}^{\omega} j._5N_{x,j}}{\sum_{j=0}^{\omega} {}_5N_{x,j}}$$

for $x = 15, 20, \ldots, 45$.

For ease of exposition of many methods, average parities in five-year age groups, 15–19, 20–24,… are often indexed as $P(i)$, $i = 1, 2\ldots$, where $P(1)$ refers to the 15–19 age group, $P(2)$ the 20–24 age group etc.

**Comparison with other estimates of average parities**

Where other fertility data are available for the same country at a roughly similar point in time, the estimates should be compared. Where the estimates diverge to any great degree, efforts should be made to understand why this might be the case, although it will often be impossible to conclude definitively which of the data sets is deficient.

**Comparison with total fertility**

As a final check, the average parity for the 45–49 age group should be compared with the estimated total fertility (TF) derived from the data on recent fertility. If fertility has been constant for a long time, and the data were accurately reported, the two measures should be very close since period and cohort fertility would be equal under these conditions. If fertility has been falling, the average parity of older women should be greater than TF. As errors of underreporting of recent fertility will artificially depress TF, while omission of older women's births will artificially depress the average parity in that group, it is important to ensure that both measures are plausible. One method of doing this uses the relational Gompertz model to examine the fertility and parity distributions and their implied relationship.

## EXAMPLE: ASSESSMENT OF DATA ON LIFETIME FERTILITY

The example below uses the data from the 2008 Census of Cambodia distributed by IPUMS. The data (weighted, to compensate for the fact that the IPUMS data represent only a microsample of the full data) are presented in Table 3.1.

The italicized cell counts represent implausible parities according to the rule-of-thumb set out earlier. The values

*TABLE 3.1 Total children ever born by age group of mother, Cambodia, 2008 Census*

| | Age group of mother | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Parity** | **15–19** | **20–24** | **25–29** | **30–34** | **35–39** | **40–44** | **45–49** | **Total** |
| 0 | 743,190 | 426,760 | 191,720 | 58,530 | 46,650 | 36,050 | 28,780 | 1,531,680 |
| 1 | 29,560 | 167,810 | 142,720 | 44,310 | 34,530 | 25,790 | 21,740 | 466,460 |
| 2 | 4,240 | 78,410 | 171,450 | 90,990 | 79,080 | 51,980 | 36,680 | 512,830 |
| 3 | 1,200 | 16,940 | 82,960 | 84,220 | 98,640 | 67,690 | 48,190 | 399,840 |
| 4 | 830 | 4,020 | 26,870 | 48,510 | 79,480 | 70,400 | 56,190 | 286,300 |
| 5 | 430 | 1,340 | 6,910 | 21,010 | 49,250 | 56,980 | 51,500 | 187,420 |
| 6 | *270* | 630 | 2150 | 8,710 | 26,020 | 37,070 | 41,420 | 116,270 |
| 7 | *120* | 380 | 630 | 3,410 | 12,530 | 23,730 | 29,680 | 70,480 |
| 8 | *80* | 200 | 400 | 1,000 | 5,450 | 12,180 | 18,320 | 37,630 |
| 9 | *60* | *100* | 120 | 350 | 2410 | 6,030 | 10,040 | 19,110 |
| 10 | *40* | *120* | 140 | 190 | 1090 | 3,120 | 5,660 | 10,360 |
| 11 | *50* | *0* | 70 | 70 | 360 | 1,420 | 2,010 | 3,980 |
| 12 | *20* | *50* | 20 | 30 | 170 | 670 | 1,350 | 2,310 |
| 13 | *10* | *10* | *0* | 10 | 60 | 270 | 410 | 770 |
| 14 | *0* | *10* | *10* | 0 | 10 | 60 | 190 | 280 |
| 15 | *0* | *0* | *10* | 0 | 20 | 90 | 150 | 270 |
| 16 | *0* | *0* | *0* | *0* | 0 | 10 | 30 | 40 |
| 17 | *0* | *0* | *0* | *0* | 0 | 10 | 30 | 40 |
| 18 | *0* | *0* | *0* | *0* | 0 | 0 | 20 | 20 |
| 19 | *0* | *0* | *0* | *0* | *0* | 0 | 10 | 10 |
| 20 | *0* | *0* | *0* | *20* | *0* | 0 | 0 | 20 |
| Unknown | 220 | 380 | 250 | 290 | 130 | 210 | 120 | 1,600 |
| **TOTAL** | **780,320** | **697,160** | **626,430** | **361,650** | **435,880** | **393,760** | **352,520** | **3,647,720** |

in these cells are summed and this total is added to the total number of women in each age group whose parity was missing in Table 3.1. The original values are then set to zero, resulting in the distribution shown in Table 3.2.

The proportion of women whose parity is unknown after making this adjustment is shown in the third last row of Table 3.2. In every age group, the proportion of women for whom parity data are missing is trivial. Although the proportion is somewhat higher in younger than in older age groups, even in the 15–19 age group only 0.11 per cent of women's parities are unknown or implausible. An el-Badry correction is therefore unnecessary and the unknown cases can be excluded from the calculation of average parities, thereby implicitly assuming that women with implausible or missing data have the same average parities as other women of the same age. (The data presented here were chosen because an el-Badry correction is not required. Chapter 4 (describing the el-Badry correction) presents data from another country whose parity data are not of as good quality.)

The proportion of women reported to be childless, shown in the second last line of Table 3.2, declines rapidly with age: by age 40, less than 10 per cent of women are still childless. As expected, this proportion falls only slightly further between the last two age groups: not many women start their childbearing after age 40. The proportion of women aged 45–49 who are childless (8.2 per cent) is relatively high. The

***TABLE 3.2*** *Total children ever born by age group of mother after correcting for implausible parities, Cambodia, 2008 Census*

| | Age group of mother | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Parity** | **15–19** | **20–24** | **25–29** | **30–34** | **35–39** | **40–44** | **45–49** | **Total** |
| 0 | 743,190 | 426,760 | 191,720 | 58,530 | 46,650 | 36,050 | 28,780 | 1,531,680 |
| 1 | 29,560 | 167,810 | 142,720 | 44,310 | 34,530 | 25,790 | 21,740 | 466,460 |
| 2 | 4,240 | 78,410 | 171,450 | 90,990 | 79,080 | 51,980 | 36,680 | 512,830 |
| 3 | 1,200 | 16,940 | 82,960 | 84,220 | 98,640 | 67,690 | 48,190 | 399,840 |
| 4 | 830 | 4,020 | 26,870 | 48,510 | 79,480 | 70,400 | 56,190 | 286,300 |
| 5 | 430 | 1,340 | 6,910 | 21,010 | 49,250 | 56,980 | 51,500 | 187,420 |
| 6 | 0 | 630 | 2150 | 8,710 | 26,020 | 37,070 | 41,420 | 116,000 |
| 7 | 0 | 380 | 630 | 3,410 | 12,530 | 23,730 | 29,680 | 70,360 |
| 8 | 0 | 200 | 400 | 1,000 | 5,450 | 12,180 | 18,320 | 37,550 |
| 9 | 0 | 0 | 120 | 350 | 2410 | 6,030 | 10,040 | 18,950 |
| 10 | 0 | 0 | 140 | 190 | 1090 | 3,120 | 5,660 | 10,200 |
| 11 | 0 | 0 | 70 | 70 | 360 | 1,420 | 2,010 | 3,930 |
| 12 | 0 | 0 | 20 | 30 | 170 | 670 | 1,350 | 2,240 |
| 13 | 0 | 0 | 0 | 10 | 60 | 270 | 410 | 750 |
| 14 | 0 | 0 | 0 | 0 | 10 | 60 | 190 | 260 |
| 15 | 0 | 0 | 0 | 0 | 20 | 90 | 150 | 260 |
| 16 | 0 | 0 | 0 | 0 | 0 | 10 | 30 | 40 |
| 17 | 0 | 0 | 0 | 0 | 0 | 10 | 30 | 40 |
| 18 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 20 |
| 19 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 10 |
| 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Unknown | 870 | 670 | 270 | 310 | 130 | 210 | 120 | 2,580 |
| **TOTAL** | **780,320** | **697,160** | **626,430** | **361,650** | **435,880** | **393,760** | **352,520** | **3,647,720** |
| Proportion missing | 0.111% | 0.096% | 0.043% | 0.086% | 0.030% | 0.053% | 0.034% | |
| Proportion childless | 95.24% | 61.21% | 30.61% | 16.18% | 10.70% | 9.16% | 8.16% | |
| Average parities | 0.0604 | 0.5833 | 1.4382 | 2.4035 | 3.1670 | 3.8126 | 4.3184 | |

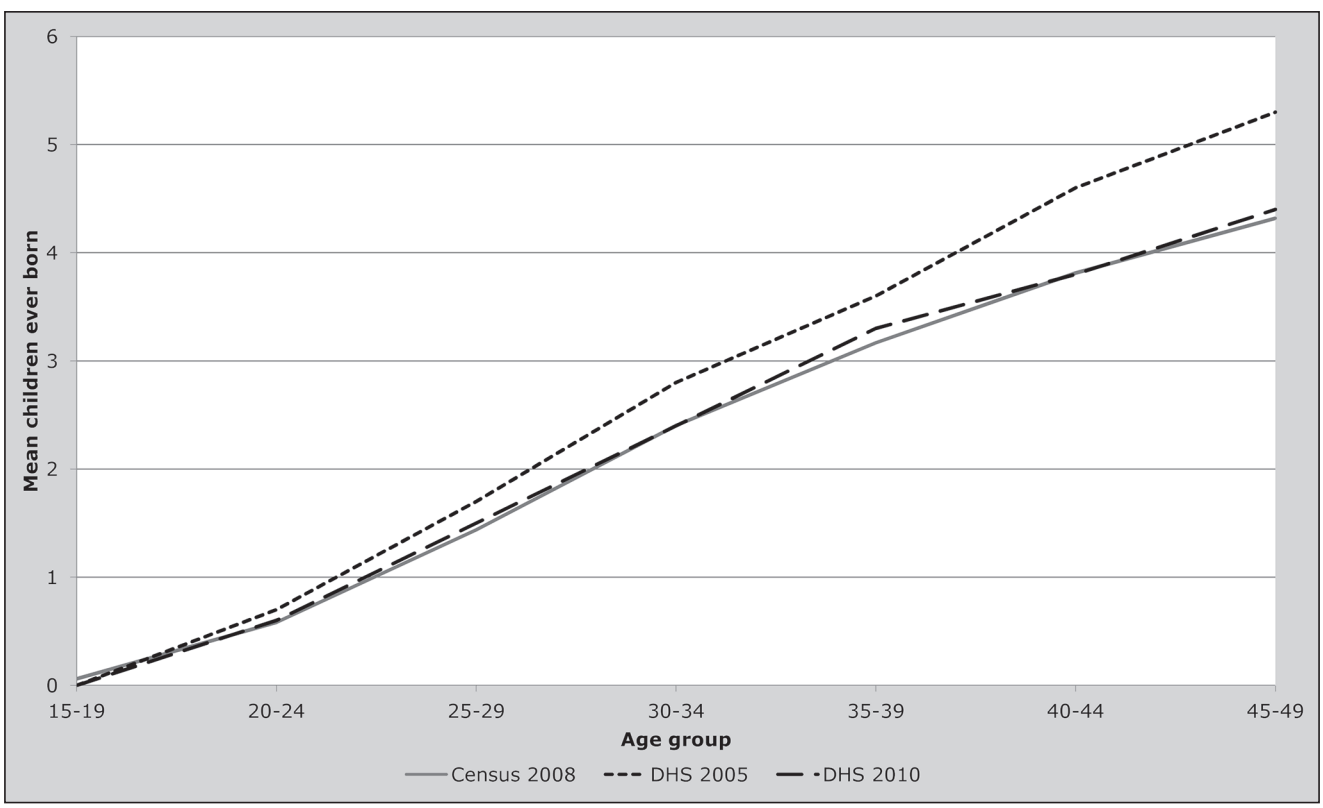
average parities suggest very low levels of fertility in teenage girls, with lifetime fertility increasing to 4.3 children per woman in the 45–49 age group. A plot of the average parities has a sigmoid shape, with the largest parity increments occurring to women in their 20s and early 30s, the ages where fertility is expected to be highest (Figure 3.1).

Figure 3.1 also shows the average parities by age group according to the 2005 and 2010 Cambodian Demographic and Health Surveys (available from the *www.statcompiler.com* DHS website). The average parities reported in the Census and the 2010 survey are very similar. However, two features suggest one should be wary of concluding that this implies that they are accurate. First, given the timing of the three enquiries, the data from the census should lie approximately half-way between the estimates from the two DHSs. This is not the case. Second, it can be seen that the average parity of women aged 40–44 in the 2005 DHS is a little higher (by 0.2 of a child) than that of women aged 45–49 in the 2010 DHS. While fertility is low among women in their late 40s in Cambodia, and random error cannot be discounted, this result should encourage a little scepticism about the data. However, overall, the average parities from the two DHSs are not fundamentally at odds with those indicated by the 2008 Census.

## REFERENCES

Feeney G. 1991. "Child survivorship estimation: Methods and data analysis", *Asian and Pacific Population Forum* **5**(2–3):51–55, 76–87. http://hdl.handle.net/10125/3600

***FIGURE 3.1*** *Average parities by age group, Cambodia, 2008 Census, 2005 DHS and 2010 DHS*



# Chapter 4 The el-Badry correction

*Tom A Moultrie*

## DESCRIPTION OF THE METHOD

The el-Badry correction is a method for correcting errors in data on children ever born caused by the enumerator or respondent failing to record answers of 'zero' to questions on lifetime fertility and, instead, leaving the response blank. When this occurs, during data processing the response is coded as 'missing' or 'unknown', even though it was evident to the enumerator at the time of data collection that the correct answer was 'zero'. The method apportions the number of women whose parity is recorded as 'missing' between those whose parity is regarded as being truly unknown, and those women who should have been recorded as childless but whose responses were left blank. It does this apportionment at an aggregate level and not on an individual basis.

## DATA REQUIRED AND ASSUMPTIONS

The method requires the number of children ever born, classified by age group of mother, including the count of women with missing data (i.e., where the field was left blank or contained an out-of-range code or a code for not answered or refused).

The method assumes that a constant proportion of women at each age truly did not state their lifetime fertility (i.e. parity) at the time of data collection. The balance of the women with unreported parities is assumed to be erroneously recorded as not stated when the women are, in fact, childless.

## CAVEATS AND WARNINGS

The method relies on the existence of a linear relationship between the proportions of women whose parity is not stated, and that of women reported to be childless. If such a linear relationship is observed, the adjusted denominator used to calculate average parities should *exclude* those women whose parity (after correction) is still regarded as unknown. This reflects the implicit assumption that these women's parity distribution is no different from those of women of the same age whose parity is known.

Where the data indicate that a correction is needed because of the large proportion of missing parity information but the method cannot be applied (for example, due to unavailability of data by age, or violation of the assumption of linearity), women of unknown parity should be *included* in the denominator used to determine average parities. This implicitly assumes that the parity of all such women is zero (i.e. that all women of unknown parity are childless). This will, of course, result in under-estimated average parities, as not all women of unknown parity are indeed childless.

## APPLICATION OF METHOD

We define $N_i = {}_5N_a$ for $a = 15, 20, \ldots, 45$ and $i = a/5 - 2$, to be the number of women in age group $i$ in the population. Thus, $N_1$ represents the number of women aged 15–19 in the population. We define $N_{i,j}$ to be the number of women in age group $i$ of parity $j$, and $N_{i,u}$ to be the number of women in age group $i$ whose parity is unknown.

***Step 1: Determine the proportion of women in each age group whose parity is a) unknown; and b) reported as zero***

Extract a table of reported children ever born ($j$) by women's age group ($i$) from the census data to obtain $N_{i,j}$. For each age group missing data on parity (i.e. blank fields and invalid codes) should be combined with codes for parity not stated to produce $N_{i,u}$. The proportion of women in age group $i$ with parity unknown is then

$$U_i = \frac{N_{i,u}}{N_i}.$$

The proportion of women in age group $i$ who are reportedly childless (i.e. are of parity zero) is given by

$$Z_i = \frac{N_{i,0}}{N_i}.$$

If the $U_i$ are small (less than 2 per cent in each age group), it is not worth applying the correction. In such a situation, average parities should be determined by assuming that the parity distribution of women with not stated parity is the same as that of women whose parity is known, by omitting the women with unstated parities from the denominator of the calculation. Thus, if $P_i$ is the average parity of women in age group $i$,

$$P_i = \frac{\sum_{j=0}^{\omega} j.N_{i,j}}{\sum_{j=0}^{\omega} N_{i,j}}.$$

If the proportions of women with parity not stated exceed 2 per cent, it is worth assessing whether the correction can be applied.

***Step 2: Plot the points ($Z_i$, $U_i$) and evaluate the data***

For the method to work correctly, the series of points ($Z_i$, $U_i$) should lie on, or very close to, a straight line. In some cases, curvature may be observed in the data points corresponding to either the oldest or the youngest ages. If the curvature affects the older ages only, even if it is quite extreme, it is acceptable to exclude the oldest, or two oldest, age groups from the fitting process and fit a straight line to the remaining points since the method has the greatest absolute impact on the proportions not stated at the youngest ages. If the curvature is most noticeable among the younger women, the method should not be used as exclusion of the data points relating to women aged 15–24 would result in the regression performing an out-of-sample extrapolation, the results of which could suggest illogical adjustments in these age groups.

If a strongly linear relationship cannot be identified, even after excluding one or two data points from older women, the method cannot be applied. In this situation, it is preferable to assume that all women of not stated parity are childless, and to include them in the denominator of the average parity calculation

$$P_i = \frac{\sum_{j=0}^{\omega} j.N_{i,j}}{N_i}. \qquad (1)$$

The analytical report should note that this has been done, and that, therefore, the average parity values are liable to be underestimated.

***Step 3: Determine the slope and intercept of the best straight line fit to the data***

The slope ($\gamma$) and intercept ($\beta$) of the fitted line are found by means of linear regression of $Z_i$ against $U_i$ applied to those data points selected for inclusion, that is, $U_i = \beta + \gamma Z_i$.

The intercept ($\beta$), which is independent of age ($i$), is the estimate of the proportion of those women in each age group with unknown parity whose parity is deemed to be truly unknown, and not misreported.

***Step 4: Estimation of the revised numbers of childless women and women whose parity is not stated***

The adjusted proportion of women in age group $i$ that is estimated to be truly childless is given by $Z_i^* = Z_i + U_i - \beta$. That is, the revised proportion of women of zero parity in any age group is the proportion actually recorded as being of zero parity together with the proportion of women in that age group of not stated parity less the estimated proportion of women whose parity is regarded as being truly unknown. The revised estimate of the number of childless women in age group $i$ is given by $N_{i,0}^* = N_i \times Z_i^*$. Thus, the estimated true proportion of women in each age group whose parity is unknown is given by $N_{i,u}^* = N_i \times \beta$. The $N_{i,j}^*$ for other parities ($j$>0) are unchanged.

***Step 5: Calculation of average parities***

If an el-Badry correction has been applied to the data, the average parities are given by

$$P_i = \frac{\sum_{j=0}^{\omega} j.N_{i,j}^*}{(1-\beta)N_i}. \qquad (2)$$

This equation embodies the assumption that the remaining women in age group $i$ of unknown parity, $\beta N_i$, who are omitted from the denominator, have the same average parity as the women in age group $i$ whose parity is known.

## INTERPRETATION AND CHECKS

The value of $\beta$ shows the estimated proportion of women whose parity is truly not stated. Larger values of $\beta$ are therefore associated with poorer quality data.

Occasionally, the method may have a contrary effect and

suggest that the number of women with not-stated parity is understated, and that the number of women of reported parity zero should be reduced. Such a situation will arise if $\beta > U_i$. If this is so, the correction should not be applied to that age group.

## WORKED EXAMPLE

The spreadsheet (see website) implements the method using data from the 1989 Kenya Census data obtained from IPUMS. The original data are presented in Table 4.1.

Inspection of the data reveals that they have been edited to disallow the recording of high parities in women aged

*TABLE 4.1 Children ever born, by age group of mother at census date, Kenya, 1989 Census*

| | Age group (*i*) | | | | | | |
|---|---|---|---|---|---|---|---|
| **Parity** | **15–19 (1)** | **20–24 (2)** | **25–29 (3)** | **30–34 (4)** | **35–39 (5)** | **40–44 (6)** | **45–49 (7)** |
| 0 | 597,560 | 198,600 | 59,400 | 23,120 | 14,580 | 11,040 | 9,560 |
| 1 | 134,700 | 224,660 | 83,140 | 26,140 | 13,620 | 9,460 | 7,740 |
| 2 | 38,120 | 202,300 | 120,940 | 38,340 | 19,180 | 13,240 | 9,280 |
| 3 | 11,120 | 126,500 | 150,500 | 53,880 | 28,020 | 17,000 | 12,440 |
| 4 | 6,820 | 59,700 | 146,500 | 73,280 | 37,340 | 21,400 | 14,800 |
| 5 | 1,740 | 33,720 | 102,300 | 87,720 | 48,140 | 28,980 | 18,560 |
| 6 | 0 | 12,480 | 58,980 | 83,580 | 56,520 | 35,260 | 26,280 |
| 7 | 0 | 0 | 57,180 | 91,800 | 56,240 | 41,260 | 28,640 |
| 8 | 0 | 0 | 0 | 64,740 | 56,560 | 42,700 | 32,920 |
| 9 | 0 | 0 | 0 | 0 | 40,780 | 39,480 | 33,000 |
| 10 | 0 | 0 | 0 | 0 | 26,840 | 32,240 | 27,920 |
| 11 | 0 | 0 | 0 | 0 | 14,920 | 22,840 | 21,920 |
| 12 | 0 | 0 | 0 | 0 | 8,280 | 14,660 | 14,720 |
| 13 | 0 | 0 | 0 | 0 | 3,740 | 7,900 | 8,920 |
| 14 | 0 | 0 | 0 | 0 | 2,180 | 4,080 | 4,900 |
| 15 | 0 | 0 | 0 | 0 | 1,260 | 2,100 | 2,860 |
| 16 | 0 | 0 | 0 | 0 | 960 | 1,200 | 1,540 |
| 17 | 0 | 0 | 0 | 0 | 520 | 680 | 1,000 |
| 18 | 0 | 0 | 0 | 0 | 420 | 520 | 620 |
| 19 | 0 | 0 | 0 | 0 | *140* | 340 | 380 |
| 20 | 0 | 0 | 0 | 0 | *160* | 300 | 280 |
| 21 | 0 | 0 | 0 | 0 | *240* | 160 | 280 |
| 22 | 0 | 0 | 0 | 0 | *40* | 100 | 60 |
| 23 | 0 | 0 | 0 | 0 | *20* | *20* | 80 |
| 24 | 0 | 0 | 0 | 0 | *60* | *20* | 80 |
| 25 | 0 | 0 | 0 | 0 | *60* | *40* | 0 |
| 26 | 0 | 0 | 0 | 0 | *60* | *40* | *80* |
| 27 | 0 | 0 | 0 | 0 | *80* | *40* | *60* |
| 28 | 0 | 0 | 0 | 0 | *20* | *40* | *40* |
| 29 | 0 | 0 | 0 | 0 | *20* | *0* | *40* |
| 30 | 0 | 0 | 0 | 0 | *340* | *440* | *360* |
| Not Stated | 402,780 | 147,540 | 61,920 | 31,580 | 20,240 | 15,420 | 12,960 |
| **TOTAL** | **1,192,840** | **1,005,500** | **840,860** | **574,180** | **451,580** | **363,000** | **292,320** |

less than 35. The editing rule applied at the preparatory stage would appear to be stricter than the one suggested in Chapter 3. Thus reports of 20–24 year old women have been restricted to parity 6 or less (rather than parity 8), reports for those aged 25–29 are truncated at parity 7 (rather than parity 12) and those of 30–34 year olds at parity 8 (rather than 15). However, implausibly high parities have been allowed to remain at ages 35 and more. Therefore, further light editing of the data highlighted in italics in Table 4.1 could be undertaken by re-assigning to the unknown category reports of parity 19 and over for age group 35–39, parity 23 and over in the age group 40–44, and parity 26 and over in the last age group, 45–49.

An option can be selected on the *Introduction* tab of the spreadsheet to set implausible parities to 'not stated' prior to the application of the method.

***TABLE 4.2*** *Correction of parity data, and calculation of proportion of women of parity zero, and parity not stated, Kenya, 1989 Census*

| | Age group (*i*) | | | | | | |
|---|---|---|---|---|---|---|---|
| **Parity** | **15–19** (1) | **20–24** (2) | **25–29** (3) | **30–34** (4) | **35–39** (5) | **40–44** (6) | **45–49** (7) |
| 0 | 597,560 | 198,600 | 59,400 | 23,120 | 14,580 | 11,040 | 9,560 |
| 1 | 134,700 | 224,660 | 83,140 | 26,140 | 13,620 | 9,460 | 7,740 |
| 2 | 38,120 | 202,300 | 120,940 | 38,340 | 19,180 | 13,240 | 9,280 |
| 3 | 11,120 | 126,500 | 150,500 | 53,880 | 28,020 | 17,000 | 12,440 |
| 4 | 6,820 | 59,700 | 146,500 | 73,280 | 37,340 | 21,400 | 14,800 |
| 5 | 1,740 | 33,720 | 102,300 | 87,720 | 48,140 | 28,980 | 18,560 |
| 6 | 0 | 12,480 | 58,980 | 83,580 | 56,520 | 35,260 | 26,280 |
| 7 | 0 | 0 | 57,180 | 91,800 | 56,240 | 41,260 | 28,640 |
| 8 | 0 | 0 | 0 | 64,740 | 56,560 | 42,700 | 32,920 |
| 9 | 0 | 0 | 0 | 0 | 40,780 | 39,480 | 33,000 |
| 10 | 0 | 0 | 0 | 0 | 26,840 | 32,240 | 27,920 |
| 11 | 0 | 0 | 0 | 0 | 14,920 | 22,840 | 21,920 |
| 12 | 0 | 0 | 0 | 0 | 8,280 | 14,660 | 14,720 |
| 13 | 0 | 0 | 0 | 0 | 3,740 | 7,900 | 8,920 |
| 14 | 0 | 0 | 0 | 0 | 2,180 | 4,080 | 4,900 |
| 15 | 0 | 0 | 0 | 0 | 1,260 | 2,100 | 2,860 |
| 16 | 0 | 0 | 0 | 0 | 960 | 1,200 | 1,540 |
| 17 | 0 | 0 | 0 | 0 | 520 | 680 | 1,000 |
| 18 | 0 | 0 | 0 | 0 | 420 | 520 | 620 |
| 19 | 0 | 0 | 0 | 0 | 0 | 340 | 380 |
| 20 | 0 | 0 | 0 | 0 | 0 | 300 | 280 |
| 21 | 0 | 0 | 0 | 0 | 0 | 160 | 280 |
| 22 | 0 | 0 | 0 | 0 | 0 | 100 | 60 |
| 23 | 0 | 0 | 0 | 0 | 0 | 0 | 80 |
| 24 | 0 | 0 | 0 | 0 | 0 | 0 | 80 |
| 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *U* | 402,780 | 147,540 | 61,920 | 31,580 | 21,480 | 16,060 | 13,540 |
| **TOTAL** | **1,192,840** | **1,005,500** | **840,860** | **574,180** | **451,580** | **363,000** | **292,320** |
| $U_i$ | 0.338 | 0.147 | 0.074 | 0.055 | 0.048 | 0.044 | 0.046 |
| $Z_i$ | 0.501 | 0.198 | 0.071 | 0.040 | 0.032 | 0.030 | 0.033 |

***Step 1: Determine the proportion of women in each age group whose parity is a) not stated; and b) equal to zero***

Table 4.2 presents the revised data, together with the calculation of the proportions of women of parity zero, and parity not stated in each age group.

The data include high proportions of women with parity not stated at ages 15–19 $\left(\frac{402,780}{1,192,840}=0.338\right)$, 20–24 (0.147) and, to a lesser extent, the older age groups. The proportion of women reported as childless ($Z_i$) falls rapidly, from around 50 per cent in the first age group down to around 3 per cent at the end of the childbearing period. On these grounds, it is worth investigating whether an el-Badry correction can be applied to the data.

***Step 2: Plot the points ($Z_i$, $U_i$) on a set of axes and evaluate the data***

The $Z_i$ and $U_i$ are plotted against each other (shown by the diamonds) in Figure 4.1. The straight line fitted to the points is shown. If a point is excluded from the fitting process, the figure in the spreadsheet represents it with an open diamond.

There is a clear linear relationship between the plotted points, and all points can be included in the application of an el-Badry correction.

***Step 3: Determine the slope and intercept of the best straight line fit***

Performing a linear regression of the $Z_i$ on the $U_i$ for the selected points gives a value for the intercept (beta) of 0.02745. This suggests that around 2.7 per cent of the data on women's parities can be regarded as truly missing.

***Step 4: Estimation of the revised numbers of childless women, and women whose parity is not stated***

The revised number of women of zero parity is given by $N^*_{i,0} = N_i\,(Z_i + U_i - \beta)$, while the revised numbers with parity unknown are calculated by multiplying the total number of women in each age group by $\beta$ as shown in Table 4.3.

For example, the number of women aged 20–24 estimated to be truly of an unknown parity is given by 0.02745× 1,005,500 = 27,603. The corrected estimate of the number of childless women aged 15–19 is derived from 1,192,840× (0.501 + 0.338 – 0.027) = 967,594.

*TABLE 4.3 Revised estimates of numbers of women with parity not stated and childless women by age, Kenya, 1989 Census*

| Age group (*i*) | Revised parity not stated | Revised zero parity |
|---|---|---|
| 15–19 (1) | 32,746 | 967,594 |
| 20–24 (2) | 27,603 | 318,537 |
| 25–29 (3) | 23,084 | 98,236 |
| 30–34 (4) | 15,763 | 38,937 |
| 35–39 (5) | 12,397 | 23,663 |
| 40–44 (6) | 9,965 | 17,135 |
| 45–49 (7) | 8,025 | 15,075 |

***Step 5: Calculation of average parities***

Since an el-Badry correction has been applied, corrected average parities, presented in Table 4.4, are then derived using Equation 2.

Note that, relative to the average parities produced if the correction is not applied (and assuming therefore that all women with not stated parity are of parity zero), the correction increases the parities in each age group by a constant, $\frac{1}{1-\beta}$.

*TABLE 4.4 Corrected average parities by age group, Kenya, 1989 Census*

| Age group (*i*) | Average parity |
|---|---|
| 15–19 (1) | 0.242 |
| 20–24 (2) | 1.525 |
| 25–29 (3) | 3.214 |
| 30–34 (4) | 4.760 |
| 35–39 (5) | 6.239 |
| 40–44 (6) | 7.120 |
| 45–49 (7) | 7.510 |

## DETAILED DESCRIPTION OF THE METHOD

The method is fully described in el-Badry (1961). el-Badry's fundamental insight was that, if it could be assumed that:

1) there is a linear relationship between the proportions of childless women of a given age in a population, and the proportion of women whose parity is not stated; and
2) the true, unknown, proportion of women whose parity is not known is a constant and independent of age, then

$$U_i = \alpha Z_i^* + \beta, \quad (3)$$

where $\alpha Z_i^*$ is the proportion of truly childless women reported as parity not stated, and $b$ is the true, constant, proportion of women with parity not stated.

Hence, if $\alpha Z_i^*$ have been misclassified as not stated when they are truly childless, then $Z_i = Z_i^* - \alpha Z_i^* = (1-\alpha) Z_i^*$. and therefore:

$$Z_i^* = \frac{Z_i}{(1-\alpha)} \quad (4)$$

and substituting this into Equation 3,

$$U_i = \frac{\alpha}{1-a} Z_i + \beta = \gamma Z_i + \beta$$

where gamma can be thought of as the odds of a childless woman being classified as being of unknown parity.

Thus, a regression of $U_i$ on $Z_i$ will give estimates of $\beta$ (as well as $\gamma$ and $\alpha$).

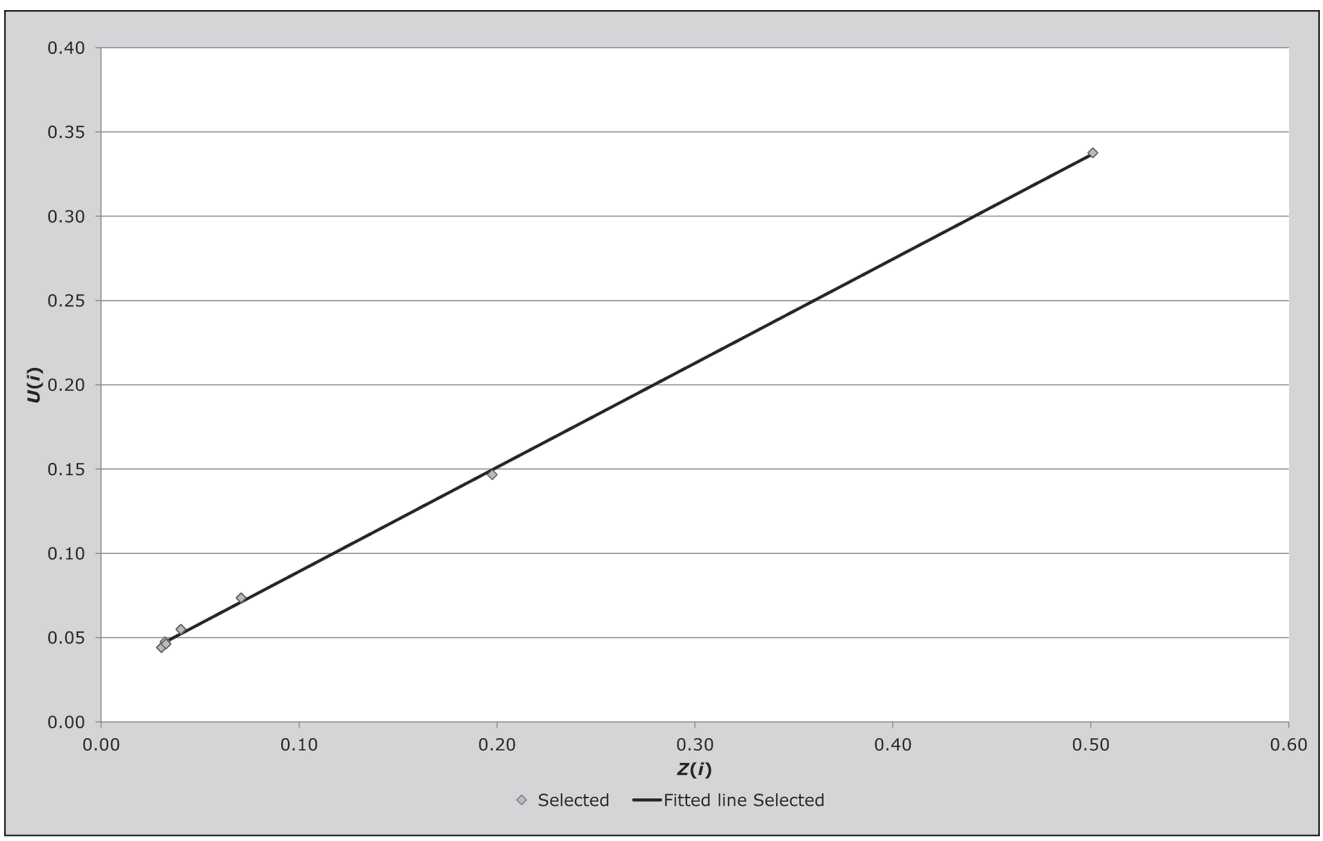
From Equation 3, we then obtain $U_i - \beta = \alpha Z_i^* = Z_i^* - Z_i$, and hence that $Z_i^* = N_{i,0}^* = U_i - \beta + Z_i$ and $U_i^* = \beta N_i$.

Note that, even though we have two identities involving $Z_i$, they will only give the same answer when the fit is exact. Convention dictates that we prefer to use Equation 3 rather than Equation 4, on the grounds that it relies on the fitted value of $\beta$ (the estimated proportion of truly not stated parities) rather than on the value of $\alpha$, which lacks intuitive interpretability.

After deriving corrected values of $Z_i^*$ and $U_i^*$, average parities can be calculated using Equation 2.

Having applied the correction, care should be taken to ensure that, in every age group, the adjusted number of childless women (that is, of parity zero) is less than the number of women reporting no births in the reference period in response to the question on recent fertility. Hence the revised $Z_i^*$ can be used to determine the minimum number of women who could not have had a birth in the reference period before the census.

***FIGURE 4.1*** *Fitting of el-Badry correction, Kenya, 1989 Census*



A version of the correction designed for (the now-rare) situations where questions on children ever born are asked only of married women is described in Annex II of *Manual X* (UN Population Division 1983).

## REFERENCES


el-Badry MA. 1961. "Failure of enumerators to make entries of zero: errors in recording childless cases in population censuses", *Journal of the American Statistical Association* **56**(296):909–924. doi: 10.1080/01621459.1961.10482134

UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

# Chapter 5 Evaluation of data on recent fertility from censuses

*Tom A Moultrie*

Before evaluating the data on recent fertility collected in a census, it is important to examine the precise wording of the questions used to capture information on recent births by consulting the questionnaire. Over successive waves of censuses, and in different countries, widely different questions have been used. The wording can influence the validity of the estimates and the direction and magnitude of biases or errors in the data.

The generic forms of the census questions on recent fertility fall into three broad categories:

- *Did you give birth in the last year (or other reference period)?*
  This question produces a simple binary answer. Multiple births in the same reference period are not captured. These could arise from the birth of twins or triplets from a single pregnancy, or from a very short birth interval separating two different pregnancies. Neither of these outcomes is likely to influence the overall fertility rate to a large extent in that birth intervals shorter than a year are rare, and the probability that a pregnancy will result in multiple births is less than 2 per cent in most settings (sub-Saharan Africa being a possible exception). When faced with data collected in this form, it is recommended that the simplifying assumption be made that all births occurred halfway through the reference period, and that only one live birth resulted from each pregnancy.
- *How many children have you given birth to in the last year (or other reference period)?*
  This question is more refined than the first form given above. It does not yield information on the timing of birth within the reference period, but it does capture information on multiple births to the same woman, without distinguishing between twins and short birth intervals. Again, it is reasonable for the purposes of calculation to assume that the births occurred halfway through the reference period.
- *What was the date of your last live birth?*
  This question seeks to identify the timing of the last delivery with a greater degree of accuracy, although typically only the month and year of the last birth are recorded. If there are follow-up questions on the number of births that occurred at that time these give more accurate information on the number of recent births.

Additional questions (for example, on the survival of the last born child; the sex of last born child; or the date of the last-but-one birth) are occasionally encountered. Answers to such questions can be used, for example, to estimate, directly from the data, child mortality rates by sex or a sex ratio at birth.

In evaluating the quality of data on recent fertility, the following checks might be conducted:

1) Comparison of the total number of births with that expected (for example, against numbers from a vital registration system, or from application of an accurate series of age-specific fertility rates to the enumerated population of women – although in the latter case, systematic underenumeration of the women might also cause the rates to be underestimated).
2) Assessment of the plausibility of the distribution of age-specific fertility rates calculated directly from the data. Plausible fertility distributions are almost invariably unimodal, concave, slightly right-skewed, and close to zero at the extremes of the childbearing age range. The distribution should also exhibit a reasonably smooth progression of fertility rates from one age to the next.
3) Plausibility checks on the reported numbers of births in the reference period. In some censuses (e.g. South Africa 1996), a significant proportion of respondents confused the questions on lifetime and recent fertility, and gave the same answers to both questions. This error manifests itself in a strong diagonal in tabulations of children ever

born by children born in the last year by age of mother (Moultrie and Timæus 2002).

4) If data on the sex of the last born child have been collected, the reported sex ratio at birth should be checked. The sex ratio at birth is usually about 1.05, but could be as low as 0.95 in African populations and up to 1.1 in some Asian populations. Values outside the range of 0.99 to 1.06 should be subjected to careful scrutiny.

In all cases, care must be taken to identify correctly the universe of women required to answer the questions (in particular the ages and marital status of eligible respondents), as well as the rules governing recording and coding of non-response and incorrect data.

**Assessment of current fertility data**

Before proceeding with an analysis of age-specific fertility rates, it is advisable to investigate the extent to which the data on recent births are missing or implausible. The absence of any missing data almost certainly indicates that the data have been edited. If this is suspected, further investigations into the extent of editing and/or imputation of the data are recommended to the extent that this is possible, for example through examination of the distribution of imputed values where imputation flag variables are included in the data.

The proportion of the data that is missing should also be checked. If this exceeds five per cent of the total number of records relevant for current fertility data, further investigations should be done. In particular, one should examine the age distribution of missing cases. If these are concentrated among young women or women in their forties this would suggest that the missing cases are missing because these mothers did not have a birth in the reference period, and no answer was recorded by the enumerator rather than an entry of zero being made. This is an error very similar to that giving rise to the el-Badry correction.

When the data are tabulated by the number of births in the reference period (as opposed to simply whether or not a birth occurred in the reference period), the distribution of single versus multiple births should be investigated. Generally, less than 2 per cent of pregnancies result in multiple births. Triplets and higher order multiple births are exceedingly rare (less than 0.5 per cent of deliveries). If the proportion of multiple births in the reference period seems too high, it is recommended that tabulations of children ever born and births in the last year are produced for each age group of women. If children ever born and births in the last year are equal in a large proportion of cases, even for parities two and over, this may suggest that respondents or enumerators did not understand the distinction between the questions on lifetime and recent fertility. However, it is possible that a large proportion of younger women with only one child ever born gave birth to that child in the reference period and a close match between lifetime reports of just one birth and recent reports of one birth in young women may not indicate reporting errors.

## DIRECT MEASUREMENT OF FERTILITY FROM CENSUS DATA

When the data are of sufficient quality, it is possible to estimate age-specific fertility rates directly. When the data are of inferior quality, age-specific fertility rates from the direct calculation are used as inputs into various methods that aim to produce more reliable estimates of the level of fertility using indirect techniques.

The exact form of the age-specific fertility rates that can be derived hinges on the nature of the data collected. An age-specific fertility rate at any given age (or in any age group) is the ratio of the number of births to women of that age (in that age group) in a defined period to the number of person-years lived by women of the same age (in the same age group) in that time period. To calculate age-specific fertility rates exactly, one would need to know reliably the exact dates of birth of mothers (to establish the mother's age) and their children. One can then calculate precisely the age of the mother at the birth of her child, as well as allocate her exposure to risk to the relevant ages or age groups over the period of investigation.

The data required for such precise calculations are not usually available in census microdata records, either because exact dates were not collected in the first place, or because of the potential for breaching confidentiality if full dates of birth are provided to end-users of the data. In addition, census data are often of insufficient quality to warrant the additional precision. Heaping of months of birth (e.g. on January) as well as years of birth (e.g. those ending in 0 or 5) are commonly encountered problems. Extended census enumeration periods can introduce problems with translating a reference period (e.g. within the last year from the interview date) to a calendar time period (e.g. 2008). Furthermore, retrospective questions about recent births asked in a census fail to capture information about births to mothers who have since died or left the country.

Four possible combinations of reporting of mother's vital information, and recent births, are typically encountered (Table 5.1).

Even in the fourth case identified in Table 5.1, which contains the most detailed information, expending effort to calculate accurately the exposure to risk for the purposes of estimating fertility is not generally warranted, as heaping of dates on particular months and other data quality problems could severely distort the resulting estimates. Thus, use of simple approximations for the calculation of fertility rates from census data is usually appropriate. Chapter 12, covering the direct measurement of fertility from survey data, describes the more precise calculation of the exposure to risk and estimation of fertility rates from data of good quality.

### *Cases 1 and 2: Estimation of age-specific fertility rates directly from the data when no information is available on the timing of the child's birth*

In the first two cases identified in Table 5.1, all that might be known about the mother's recent fertility is whether or not she gave birth to at least one child in the period before the census. In more informative variants of the recent fertility question, the mother may be asked about the number of live births in the period preceding the census. Such a question allows the identification both of multiple births from the same pregnancy (twins, triplets etc.), as well as instances of more than one pregnancy ending in the defined period.

Since the mother's age at birth is not known, the approximation usually used is to tabulate the fertility rates by the reported age of the mother at the census date. The additional assumption is then made that all births occurred half-way through the interval in question. This means that mothers are, on average, older by half the interval length at the time of the census, with the implication that the ages to which the fertility rates actually refer are younger than the reported ages at census. Most standard methods of estimating fertility indirectly compensate for the displacement of ages arising from this mismatch.

The additional information (on mother's month and year of birth) available in the second case is not particularly helpful in refining the estimates of fertility since additional assumptions of uniformity of the distribution of children's birthdays are still required. Thus, where the data that were collected fall into either the first or the second case identified in Table 5.1, fertility rates are estimated by dividing the count of children reported born in the reference period (by age of mother at the census date) by the number of women of that age. The total number of births in the reference period reported by women aged $x$ at the census date, $B_x$, is given by

$$B_x = \sum_{k=0}^{\omega} k.N_{x,k}$$

where $k$ is the reported number of births in the reference period, $\omega$ is the maximum value of $k$ in the data and $N_{x,k}$ is the number of women aged $x$ at the census reporting $k$ births in the last year. If $\omega$ is classified as an open interval, e.g. 3+ births in the reference period, women in that category are all assumed to have had the number of births that opens that interval. Again, the error thus introduced is small.

The number of women aged $x$ is given by

$$N_x = \sum_{k=0}^{\omega} N_{x,k}.$$

Women whose recent births are unknown or unrecorded must be excluded from both the numerator and denominator, with the implicit assumption that their fertility is no different from that of women whose recent fertility is known. Age-specific fertility rates (ASFRs) at age $x$ are given by

*TABLE 5.1 Taxonomy of data on mother and children for estimating recent fertility*

| | Mothers' vital information | |
|---|---|---|
| Reporting of children born in the preceding period[†] | Age in completed years at census | Date of birth (at least month and year) |
| Number of children born (or simple binary, yes/no) | (1) | (2) |
| Date of birth of last born child (at least month and year) | (3) | (4) |

[†]Typically the preceding period is 12 months, but analysts should be alert to non-standard reference periods, for example based on time elapsed since an important national event or holiday

$$f_x = B_x / N_x.$$

Using the conventional age range (from 15 to 49, inclusive) as the limits for the summation, the implied Total Fertility (TF) from the single-age data is

$$TF = \sum_{a=15}^{49} f_a.$$

Total fertility is a synthetic cohort measure – indicating the number of children a woman would have if she survives to age 50 (deemed to be the end of childbearing) and experiences the age-specific fertility rates currently observed immediately before the census throughout her reproductive life.

Fertility rates by single years of age should be calculated and plotted to check the internal coherence of the data. The ASFRs will tend to be less erratic than either the numerators or the denominators on their own, and may indicate plausible levels and distributions of fertility. A highly erratic series of age-specific fertility rates by age, departing markedly from the anticipated *n*-shape, offers a strong indication that the recent fertility data are problematic, and suggests that further investigations are required.

Finally, age-specific fertility rates in conventional five-year bands, ${}_5 f_x$, where $x = 15, 20, \ldots, 45$, can be derived:

$$f_i = {}_5 f_x = \frac{\sum_{x=5i+10}^{5i+14} B_x}{\sum_{x=5i+10}^{5i+14} N_x}$$

where the index, $i$, is determined by the relation $i = (x/5) - 2$. The measure of total fertility is thus

$$TF = 5 \sum_{i=1}^{7} f_i.$$

While the TF is an age-standardized measure of fertility (implicitly assuming a uniform distribution of the population of child-bearing population of women by age in each age group), the fertility rate in any age group is not standardized within the group. As a result, the TF derived from calculations using age-groups and single years of age will differ to a small degree, typically in the second or third decimal place.

Total fertility should be compared with estimates from other data sources from the same country (e.g. DHS). It is worth remembering, however, that the ASFRs and TF produced using this method do not take into account the true exposure-to risk in the derivation of the denominator. In addition, the numerator includes events that took place during the reference period categorized by the age of the mother at the end of the reference period, not by her age at the time the event took place. Most methods of indirect fertility estimation adjust the derived fertility rates to account for this age shift. For purposes of basic comparison (that is, assessing the shape and level of the fertility distributions), the differences in classification by age are not of major importance. However, the *F*-only variant of the relational Gompertz model provides a method of unshifting fertility rates while smoothing them, should this be desired.

## EXAMPLE: DIRECT CALCULATION OF FERTILITY

In the 2008 Cambodian Census, women were asked about the number of children they gave birth to in the previous year. Mother's age was classified by age at the census date. The data are shown in Table 5.2.

The 'missing' column shows that only 1,600 women, out of nearly 3.65 million aged between 15 and 49, did not have their recent fertility recorded. This represents 0.04 per cent of all women, and will have no material impact on the estimated fertility of women in Cambodia. A further check on the age distribution of these cases shows no clear age pattern of omission. The number of births is given by the weighted sum of women reporting 1, 2, 3 and 4 deliveries, in the last row. This calculation shows that 173,070 women (170,910 + 1,760 + 250 + 150) gave birth to a total of 175,780 births (1×170,910 + 2×1,760 + 3×250 + 4×150) during the year preceding the census. Of these women, 98.8 per cent (170,910/173,070) experienced a single birth. 1.0 per cent had twins, and 0.2 per cent triplets or higher-order multiple births. The possibility of quintuplets (or five births in two deliveries over the period) is remote and need not be considered. Had the census not counted the multiple births separately, the crude birth rate would have been under-estimated by a factor of 173,070/175,780 = 0.984. This represents an under-estimate of just 1.6 per cent.

Using the data above, the series of single-age ASFRs is derived by dividing the total number of births to women of each age by the number of women reporting their current fertility, that is, excluding those women who did not report how many births they had in the last year. The rates are shown in Figure 5.1. Even though the number of women

*TABLE 5.2 Recent fertility by age of mother at the census date, Cambodia, 2008 Census*

| | Births in the last year | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Age | 0 | 1 | 2 | 3 | 4 | Missing | Births | Women | ASFR |
| 15 | 160,980 | 120 | 0 | 0 | 0 | 80 | 120 | 161,180 | 0.0007 |
| 16 | 152,710 | 500 | 0 | 0 | 0 | 50 | 500 | 153,260 | 0.0033 |
| 17 | 144,970 | 1,250 | 10 | 10 | 0 | 20 | 1,300 | 146,260 | 0.0089 |
| 18 | 182,500 | 3,540 | 20 | 0 | 0 | 40 | 3,580 | 186,100 | 0.0192 |
| 19 | 127,840 | 5,640 | 10 | 0 | 0 | 30 | 5,660 | 133,520 | 0.0424 |
| 20 | 147,990 | 8,840 | 80 | 0 | 0 | 90 | 9,000 | 157,000 | 0.0574 |
| 21 | 123,960 | 9,500 | 30 | 0 | 0 | 70 | 9,560 | 133,560 | 0.0716 |
| 22 | 126,030 | 11,600 | 80 | 0 | 0 | 30 | 11,760 | 137,740 | 0.0854 |
| 23 | 123,750 | 11,830 | 70 | 10 | 0 | 110 | 12,000 | 135,770 | 0.0885 |
| 24 | 121,820 | 11,010 | 150 | 10 | 20 | 80 | 11,420 | 133,090 | 0.0859 |
| 25 | 137,460 | 12,420 | 100 | 0 | 0 | 60 | 12,620 | 150,040 | 0.0841 |
| 26 | 115,370 | 11,320 | 110 | 0 | 0 | 80 | 11,540 | 126,880 | 0.0910 |
| 27 | 117,840 | 11,580 | 190 | 0 | 0 | 40 | 11,960 | 129,650 | 0.0923 |
| 28 | 118,270 | 10,690 | 110 | 0 | 10 | 30 | 10,950 | 129,110 | 0.0848 |
| 29 | 82,990 | 7,600 | 120 | 0 | 0 | 40 | 7,840 | 90,750 | 0.0864 |
| 30 | 77,690 | 5,950 | 40 | 10 | 0 | 30 | 6,060 | 83,720 | 0.0724 |
| 31 | 58,800 | 4,820 | 50 | 20 | 0 | 30 | 4,980 | 63,720 | 0.0782 |
| 32 | 67,110 | 4,480 | 150 | 20 | 0 | 110 | 4,840 | 71,870 | 0.0674 |
| 33 | 67,080 | 4,240 | 40 | 0 | 0 | 50 | 4,320 | 71,410 | 0.0605 |
| 34 | 67,010 | 3,800 | 30 | 10 | 10 | 70 | 3,930 | 70,930 | 0.0555 |
| 35 | 90,720 | 4,570 | 60 | 20 | 0 | 30 | 4,750 | 95,400 | 0.0498 |
| 36 | 77,950 | 3,800 | 10 | 10 | 0 | 30 | 3,850 | 81,800 | 0.0471 |
| 37 | 81,320 | 4,070 | 50 | 10 | 10 | 10 | 4,240 | 85,470 | 0.0496 |
| 38 | 92,290 | 3,780 | 30 | 20 | 30 | 30 | 4,020 | 96,180 | 0.0418 |
| 39 | 74,030 | 2,920 | 50 | 0 | 0 | 30 | 3,020 | 77,030 | 0.0392 |
| 40 | 88,940 | 2,720 | 70 | 10 | 10 | 50 | 2,930 | 91,800 | 0.0319 |
| 41 | 71,250 | 2,140 | 0 | 0 | 0 | 20 | 2,140 | 73,410 | 0.0292 |
| 42 | 81,560 | 2,010 | 30 | 0 | 0 | 60 | 2,070 | 83,660 | 0.0248 |
| 43 | 72,930 | 1,270 | 10 | 0 | 0 | 30 | 1,290 | 74,240 | 0.0174 |
| 44 | 69,660 | 930 | 10 | 0 | 0 | 50 | 950 | 70,650 | 0.0135 |
| 45 | 84,290 | 760 | 30 | 10 | 10 | 30 | 890 | 85,130 | 0.0105 |
| 46 | 67,330 | 510 | 0 | 50 | 30 | 40 | 780 | 67,960 | 0.0115 |
| 47 | 66,220 | 270 | 10 | 0 | 10 | 0 | 330 | 66,510 | 0.0050 |
| 48 | 74,790 | 310 | 10 | 10 | 0 | 30 | 360 | 75,150 | 0.0048 |
| 49 | 57,600 | 120 | 0 | 20 | 10 | 20 | 220 | 57,770 | 0.0038 |
| **TOTAL** | **3,473,050** | **170,910** | **1,760** | **250** | **150** | **1,600** | **175,780** | **3,647,720** | **1.6157** |

enumerated at each age is erratic, the ASFRs by single years of age are relatively smooth, with a clearly defined fertility pattern and a typical peak in the mid-twenties.

According to these data, total fertility is 1.61 children per woman. Summing births and women in five-year age groups produces the same answer (Table 5.3), although, as suggested above, the measures do differ in the third decimal place.

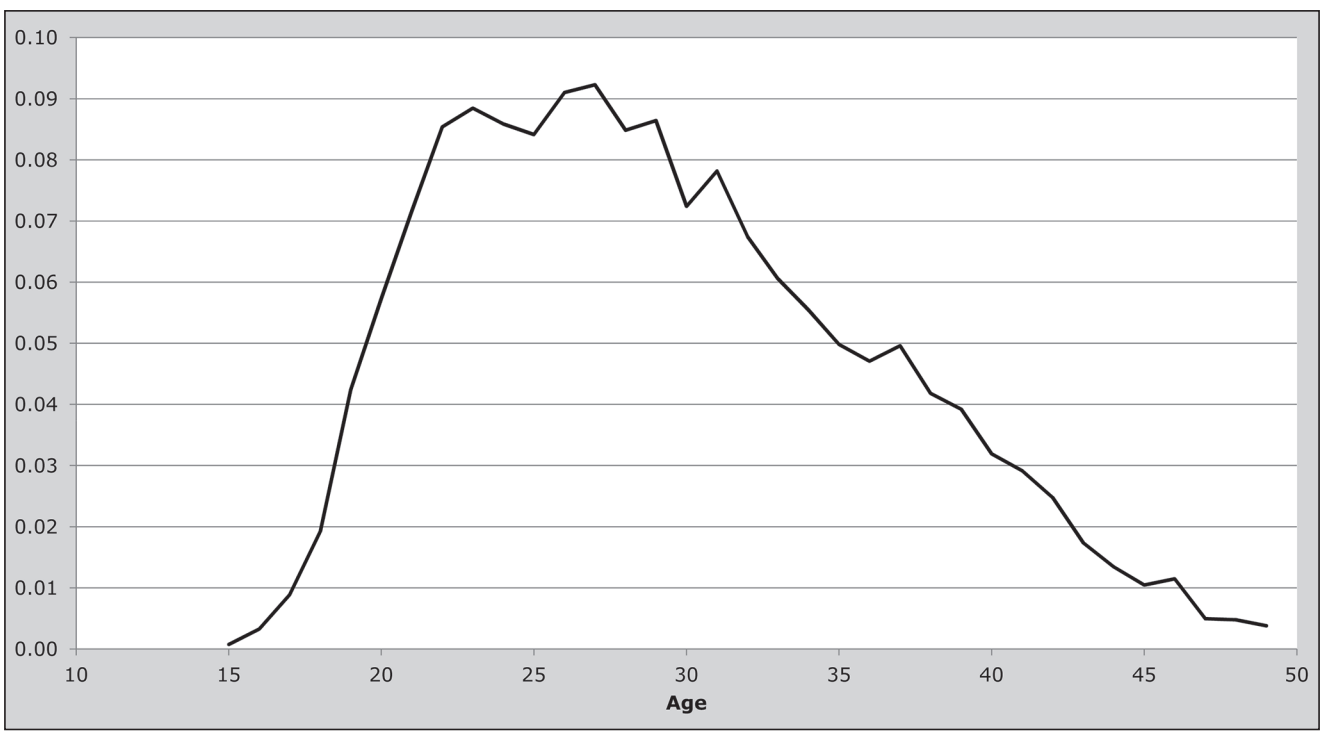
Even in the absence of external checks, the results from the 2008 Census data suggest implausibly low levels of fertility in Cambodia. The data are also inconsistent with the average parities calculated in Chapter 3. This suggests

***TABLE 5.3*** *Age-specific fertility rates in five-year age groups, Cambodia, 2008 Census and 2005 and 2010 Demographic Health Surveys*

| Age group | Women | Missing | Births | ASFR | DHS2005 | DHS2010 |
|---|---|---|---|---|---|---|
| 15–19 | 780,320 | 220 | 11,160 | 0.014 | 0.047 | 0.046 |
| 20–24 | 697,160 | 380 | 53,740 | 0.077 | 0.175 | 0.173 |
| 25–29 | 626,430 | 250 | 54,910 | 0.088 | 0.180 | 0.167 |
| 30–34 | 361,650 | 290 | 24,130 | 0.067 | 0.142 | 0.121 |
| 35–39 | 435,880 | 130 | 19,880 | 0.046 | 0.091 | 0.071 |
| 40–44 | 393,760 | 210 | 9,380 | 0.024 | 0.041 | 0.028 |
| 45–49 | 352,520 | 120 | 2,580 | 0.007 | 0.005 | 0.004 |
| TF | | | | 1.61 | 3.41 | 3.05 |

Source: Census estimates, own calculations; DHS StatCompiler (*www.statcompiler.com*)

***FIGURE 5.1*** *Age-specific fertility rates, Cambodia, 2008 Census*



that the data on recent fertility collected in this census are seriously deficient. This is confirmed by external checks, in the form of estimates of fertility from two DHSs conducted before and after the census. The data in the last two columns of Table 5.3 show that the estimate of total fertility in the 2010 DHS (based on births in the three years before the survey) was 3.1 children per woman. The estimate of total fertility from the 2005 DHS was 3.4 children per woman. It appears that only about half the births that occurred in the year before the census were reported to census enumerators.

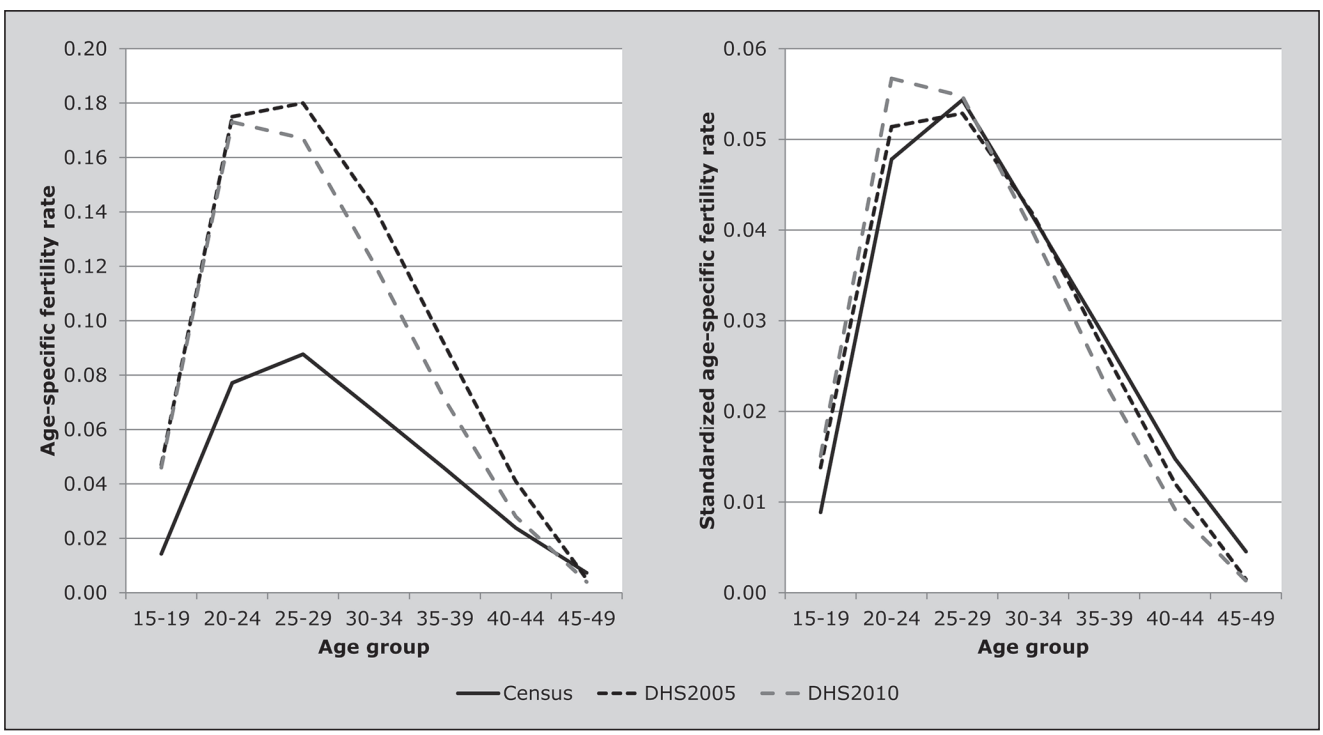
The left-hand panel of Figure 5.2 shows the age-specific fertility rates calculated from the 2008 Census and the two DHSs. Clearly the fertility rates implied by the census are out of line relative to the DHSs. The latter in turn, show a rather strange pattern of fertility change over the five years, driven by almost constant reductions in fertility between ages 25 and 44. The right-hand panel of Figure 5.2 shows the same rates, but this time standardized to a TF of one child per woman. Despite substantial differences in the implied level of fertility, the shape of the three fertility distributions are similar, with the only real difference between them being in the 2024 age group. It is unlikely, therefore, that there were significant differentials in the quality of the reporting of recent fertility in the 2008 Cambodia Census according to the age of women.

This result suggests that, even though the level of fertility implied by the 2008 Census data is seriously flawed, the shape of the fertility distribution is reasonably accurate. This is a prerequisite for applying many of the indirect methods of fertility estimation.

### *Cases 3 and 4: Estimation of age-specific fertility rates when information is available on the timing of the child's birth*

If the births are classified by women's date of last birth, a suitable period for the fertility investigation needs to be chosen. In general, it is advisable not to use a period much longer than a year as longer periods of investigation increase the probability that women might have had more than one

*FIGURE 5.2 Age-specific fertility rates, and standardized age-specific fertility rates, Cambodia, 2008 Census, 2005 DHS and 2010 DHS*



pregnancy in that period. This results in births earlier in time being omitted (the requirement being to report on the date of birth of the last child, not all children in the period), meaning that estimates of fertility will systematically exclude births in the more distant past. In addition, if fertility has been changing rapidly, extending the period of investigation over more than a year means that the resulting estimates represent some kind of average of fertility over the period. If the census was conducted fairly early or late in the year, however, there is potentially some advantage to basing the rates on births since the beginning of the previous or current year respectively as this does not require women to remember the month of birth of their child accurately. The number of births reported in the reference period can then be prorated to produce an estimate of annual births. Rates can be calculated both in this way and based on a 12-month reference period and the results compared.

The third scenario in Table 5.1 does not permit the derivation of a completely accurate measure of fertility, as the age of the mother at the birth of the child cannot be established precisely. However, knowledge of the child's date of birth does permit the numerator of the age-specific fertility rates to be derived more carefully.

In the commonly-encountered situation where the question asked is about the month and year of the last child's birth, a more careful approach can be taken to determining the number of births in the last year. Usually a notional census date is defined. The questions on the census questionnaire typically refer to a particular day, even if the actual process of enumeration takes several weeks. A list of census dates for the last three rounds of censuses is maintained by the UN at *http://unstats.un.org/unsd/demographic/sources/census/censusdates.htm*; a list of census dates for data maintained by IPUMS is available at *https://international.ipums.org/international/samples.shtml*.

In establishing the numerator, all the births reported in the month of the census, and a prorated proportion of births that are reported to have occurred in the equivalent month a year earlier should be included. To extract this information from census data, the date handling capacity of the statistical package being used, or the DHS Century-Month Code (CMC) system can be used.

*TABLE 5.4 Births reported in each month by age of mother at census date (24–25 August 1999), Kenya, 1999 Census*

| | Age of mother at census | | | | | | |
|---|---|---|---|---|---|---|---|
| **Month** | **15–19** | **20–24** | **25–29** | **30–34** | **35–39** | **40–44** | **45–49** |
| August 1998 | 13,240 | 31,300 | 23,120 | 13,940 | 8,940 | 3,220 | 560 |
| September 1998 | 9,800 | 22,900 | 17,260 | 9,560 | 6,180 | 2,080 | 680 |
| October 1998 | 9,240 | 21,580 | 15,520 | 9,600 | 5,880 | 1,880 | 500 |
| November 1998 | 9,040 | 21,940 | 16,060 | 9,880 | 5,280 | 1,660 | 540 |
| December 1998 | 10,200 | 23,700 | 18,000 | 10,580 | 5,940 | 2,080 | 480 |
| January 1999 | 14,720 | 28,620 | 20,620 | 12,260 | 7,300 | 2,180 | 660 |
| February 1999 | 20,740 | 42,140 | 30,860 | 17,400 | 11,220 | 4,560 | 2,060 |
| March 1999 | 15,620 | 31,480 | 21,320 | 12,520 | 7,340 | 2,820 | 520 |
| April 1999 | 18,660 | 33,160 | 24,260 | 12,240 | 7,820 | 2,860 | 720 |
| May 1999 | 19,660 | 33,880 | 22,860 | 13,960 | 7,440 | 2,480 | 760 |
| June 1999 | 20,100 | 32,140 | 23,380 | 12,580 | 7,300 | 2,720 | 560 |
| July 1999 | 21,600 | 32,360 | 23,860 | 13,800 | 7,060 | 2,640 | 520 |
| August 1999 | 15,900 | 25,020 | 16,720 | 9,280 | 5,840 | 1,620 | 360 |
| Estimated births in the last year | 188,269.68 | 355,987.74 | 255,940.65 | 146,807.74 | 86,618.71 | 30,307.10 | 8,486.45 |
| Number of women | 1,700,060 | 1,495,180 | 1,205,060 | 849,620 | 725,780 | 519,740 | 417,500 |
| Age-specific fertility rates | 0.1107 | 0.2381 | 0.2124 | 0.1728 | 0.1193 | 0.0583 | 0.0203 |

In the 1999 Kenya Census, the official census date was the night of 24–25 August 1999. To estimate the births that occurred in the year preceding the census, all births reported between September 1998 and August 1999 would be included, along with 1–24/31 (=7/31) of the births reported in August 1998. This assumes that births are uniformly distributed over the days of a month (Table 5.4).

The estimated number of births in the year before the census in the 30–34 age group, for example, is then given by

$$\frac{7}{31}(13,940)+9,560+9,600+\ldots+9,280=146,807.74\ .$$

In the absence of further information about the mother's date of birth, the data above are tabulated according to the mother's age at the census date. As noted above, the rates so derived would thus be subject to a half-year shift.

Dividing these births by the number of women in each age group gives the age-specific fertility rates. The resulting estimate of total fertility of 4.66 children per woman is clearly out of line with other estimates of fertility in the country for around that time. This, as with Cambodia, suggests that widespread underreporting occurred of births reported in the year before that census.

Only in the fourth case, when detailed information is available on both mother and child's date of birth, is it possible to produce a precise measurement of fertility. However, if there is evidence of extreme heaping of reported dates of birth (for example on 1 January), there is little point in making use of the more refined measures as they will be distorted by the heaping. Thus, since the quality and internal consistency of the data collected in a census are unlikely to be as good as in a DHS, it is inappropriate to attempt the precise calculation of fertility rates that one would with a DHS. In some situations, however, the extent of heaping in the reported dates of birth and other errors in the data may be sufficiently limited to merit calculation of direct estimates of fertility. In these situations, the principles outlined for the calculation of estimates of recent fertility from survey data should be applied.

## REFERENCES

Moultrie, TA and IM Timæus. 2002. *Trends in South African Fertility between 1970 and 1998: An Analysis of the 1996 Census and the 1998 Demographic and Health Survey.* Cape Town: Medical Research Council. http://www.mrc.ac.za/bod/trends.pdf

# One Census Methods

# Chapter 6 Overview of fertility estimation methods based on the *P*/*F* ratio

*Tom A Moultrie*

Almost all methods of estimating fertility indirectly have their origins in the *P*/*F* ratio method first proposed by Brass (1964). In addition, the interpretation of the results from other methods (for example, cohort-period fertility rates) and some of the diagnostic tools used to assess the quality of the data when estimating child mortality also rely on the intrinsic logic of the *P*/*F* ratio approach. Thus, while the method in its original and modified forms has been superseded by the relational Gompertz model and its variants, it is useful to present the essential logic of the method here. The interested reader is referred to *Manual X* (UN Population Division 1983) for a full exposition of the approach.

## THE BRASS *P/F* RATIO METHOD

The foundation of the method rests on the observation that if fertility has been constant for an extended period of time, cohort and period measures of fertility will be identical. In other words, under conditions of constant fertility, the cumulated fertility of a cohort of women up to any given age will be the same as the cumulated fertility up to that same age in any given period.

If we assume that there are no appreciable mortality differentials by the fertility of mother, so that surviving women do not have materially different levels of childbearing from deceased women, the cumulated fertility of a cohort of women up to any given age is the same as the average parity in that cohort. (This assumption is not very important as even if there are differentials in the fertility of living and deceased women, in most populations the magnitude of female mortality in the reproductive ages is very small and the effect of differential survival will therefore be small.)

Brass defined *P* to be the average parity (cumulated lifetime fertility) of a cohort of women up to a given age, and *F* to be closely related to the cumulated current (period) fertility up to that same age. The *P*/*F* ratio method expresses these two quantities in relation to each other in the form of a ratio for each age group.

The derivation of *F* is a little more complicated than suggested above for two reasons. First, any comparison of cohort and period fertility has to deal with the probable shifting of the data on recent fertility brought about by the question being based on the age of the mother at the time of the inquiry rather than her age at the time of her most recent birth. Second, while the cumulation of period fertility to any given age will reflect the fertility experience of all women up until that age, the average parities typically calculated reflect those of women in 5-year age groups and hence reflect (approximately) the average parity of women aged at the midpoint of that age group. The method formulated by Brass addresses both these aspects.

It follows that if fertility has been constant in a population for an extended period of time, and if the data are free of error, the *P*/*F* ratio would equal 1 in every age group. If fertility has been falling, however, cumulated life time fertility would be greater than cumulated current fertility. In this case (in the absence of errors in the data) the *P*/*F* ratio would depart from unity systematically with increasing age of mother.

The corollary to this observation is that one would expect the *P*/*F* ratio to be fairly close to unity at the youngest ages because even by women's mid-twenties one would not expect significant deviation of cumulated period fertility from cumulated lifetime cohort fertility as most of the births to women in that cohort would have happened fairly recently. It is from this observation that the *P*/*F* ratio derived from women aged 20–24 at the time of a survey is held to be the most reliable indicator of the quality of the fertility data collected. Conveniently, the supposition is that the average

parities of younger women are usually fairly accurately reported, at least relative to those of older women.

It is this characteristic pattern of departure from unity with age of mother that forms the basis for many diagnostic investigations into the nature and quality of data drawn from questions based on recent and lifetime fertility.

## DIAGNOSTICS BASED ON THE *P/F* RATIO

In reality the data are never free from error, and so the hypothetical pattern of departure of the *P*/*F* ratio from unity is confounded and obfuscated by underlying errors in the data.

As discussed in Chapters 3 and 4, two errors typically affect these data. The first is that reports on lifetime fertility – that is, cumulated cohort fertility – become increasingly inaccurate with age of the respondent, with older women tending to under-report their lifetime fertility. Errors of this kind will therefore tend to depress the numerator of the *P*/*F* ratio, particularly at the older ages. If such errors occur in the data, the ratio will tend to be closer to unity than it might truly be.

The second kind of error frequently encountered is that women tend to under-report recent births, regardless of their age. Errors of this type will result in the reported level of recent fertility being somewhat lower than anticipated, thereby causing the *P*/*F* ratio to be inflated.

The *P*/*F* ratio method seeks to correct the second problem by applying the *P*/*F* ratio applicable to younger women (for the reasons set out above) to the directly observed fertility schedule as a scaling factor.

## SUMMARY OF METHODS BASED ON THE *P/F* RATIO METHOD

A number of methods described here were originally presented in *Manual X* as extensions of the *P*/*F* ratio method. The relational Gompertz model can be thought of as an improved and more versatile version of the Brass *P*/*F* ratio method. The model uses the same input data (and makes the same assumptions about errors that affect fertility data) as its precursor. Importantly, however, the method does not require an assumption that fertility has been constant in the past. Nonetheless, the comparison of lifetime and period fertility lies at the heart of the method.

Most of the extensions to the Brass *P*/*F* ratio method presented in *Manual X* have been recast as extensions to the relational Gompertz model. These extensions include those methods that make use of the data on parity increments from two censuses to estimate fertility; methods that use parity increments in conjunction with a schedule of intercensal fertility rates (the synthetic relational Gompertz model); and indirect methods that make use of data from vital registration systems. Cohort-period fertility rates derived from survey data also rely on the logic of the *P*/*F* ratio method to shed light on longer-term trends and dynamics in fertility.

## REFERENCES


Brass W. 1964. *Uses of census or survey data for the estimation of vital rates*. Paper prepared for the African Seminar on Vital Statistics, Addis Ababa 14–19 December 1964. Document No. E/CN.14/CAS.4/V57. New York: United Nations.

UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

# Chapter 7 The relational Gompertz model

*Tom A Moultrie*

## DESCRIPTION OF METHOD

The relational Gompertz method is a refinement of the Brass *P/F* ratio method that seeks to estimate age-specific and total fertility by determining the shape of the fertility schedule from data on recent births reported in censuses or surveys while determining its level from the reported average parities of younger women.

In producing estimates of age-specific and total fertility, the method seeks to remedy the errors commonly found in fertility data associated with too few or too many births being reported in the reference period, and the under-reporting of lifetime fertility and errors of age reporting among older women. These errors are described in greater detail in the chapters on evaluation of fertility data (Chapters 3 and 5).

The method relies on a useful property of a (cumulated) Gompertz distribution, $G(x) = \exp\left(a.\exp\left(bx\right)\right)$, which is sigmoidal (i.e. S-shaped), but also has an associated hazard function that is right-skewed and which therefore captures fairly well both the pattern of average parities of women by age and their cumulated fertility. The form of $G(x)$ implies that a double-negative log transform of proportional cumulated fertilities or average parities approximates a straight line for most of the age range. The double-log transform, $Y(x) = -\ln\left(-\ln\left(G(x)\right)\right)$, is termed a *gompit* and has a close analogue in the logit transform frequently used in mortality analysis. Brass, however, found that a much closer linear fit could be obtained by a relational model that expresses the gompits of an observed series of fertility data as a linear function of the gompits of a defined standard fertility schedule. In other words, $Y(x) = \alpha + \beta Y^s(x)$, where $Y^s(x)$ is the gompit of the standard fertility schedule. Evidently, if $\alpha = 0$ and $\beta = 1$, the fertility schedule will be identical to the standard fertility schedule. Alpha ($\alpha$) represents the extent to which the age location of childbearing in the population differs from that of the standard (negative values imply an older distribution of ages at childbearing than in the standard), while beta ($\beta$) is a measure the spread of the fertility distribution (values greater than 1 imply a narrower distribution).

As input data, the method requires average parities at each age group, ${}_5P_x$, for $x = 15,20,\ldots, 45$, and fertility rates in each age group, ${}_5f_x$. For ease of exposition, and to differentiate more clearly between lifetime and recent fertility data, ${}_5P_{15}$ is indexed as $P(1)$, ${}_5P_{20}$ as $P(2)$ and so on. The derivation of these inputs from census data is described in Chapter 3. As with other methods, the average parities should be adjusted with an el-Badry correction where appropriate.

Cumulated (period) fertility to the end-point of each age group, $F(x)$ is given by

$$F(x+5) = 5.\sum_{a=15,5}^{x} {}_5f_a \ .$$

The original method proposed by Brass (1978) used the series of the gompits of the ratio of cumulated fertility to the end of each age group to the fertility rate cumulated to age 50 (i.e. total fertility, TF), giving a sigmoidal curve with minimum of 0 and a maximum (at the last age group) of 1. Gompits of the average parities are derived in a similar manner.

There are two inherent weaknesses in this approach. First, it requires total fertility as an input, and estimates of total fertility available from reported age-specific fertility rates (ASFRs) may be biased. In fact, total fertility is often the parameter of greatest interest that the analyst is trying to estimate. The second weakness is the implicit assumption of constant fertility over time arising from the treatment of the parity gompits. Nonetheless, Brass' formulation inspired the derivation of the standard fertility schedule by Booth (1980, 1984), which is still used in the model to this day.

Both limitations are addressed comprehensively by Zaba's (1981) reformulation of the method, which avoids the circularity of the original method while also dropping

the need to assume that fertility has been constant. Further unpublished work by Zaba generalized the approach to incorporate alternative variants of the model (some of which are described here). A full exposition of Zaba's reformulation is given in a subsequent section. In summary, however, she showed that the model can be expressed as

$$z(x) - e(x) = \alpha + \beta g(x) + \frac{c}{2}(\beta - 1)^2, \qquad (1)$$

where $e(x)$, $g(x)$ and $c$ are functions of the chosen standard and $z(x)$ is the gompit of the ratios of adjacent cumulated period fertility measures, i.e. $F(x)/F(x+5)$, instead of $F(x)/50$ as Brass originally suggested. In other words,

$$z(x) = -\ln\left(-\ln\left(\frac{F(x)}{F(x+5)}\right)\right).$$

For the parity data, the model is fitted to the ratios of adjacent average parities, $P(i)/P(i+1)$. This means that the model can be used without the need to estimate total fertility before fitting the shape parameters. It follows further from Equation 1 that a plot of $z(x) - e(x)$ against $g(x)$ should be a straight line with slope $\beta$ and intercept $\alpha + \frac{1}{2}c(\beta - 1)^2$. (Noting that $\beta$ should be close to one, early formulations of the procedure deemed the last term of the intercept unimportant, leaving the intercept approximated by alpha. With the computing power now to hand, there is no justification for the associated loss of precision in the calculation of the intercept. However, the requirement that $\beta$ be close to 1 remains).

Exactly the same reasoning holds for the evaluation of the parity data. Using $P(i)/P(i+1)$, the ratio of average parities in successive age groups, with a linear equation relating $z(i)$–$e(i)$ to $g(i)$ results in

$$z(i) - e(i) = \alpha + \beta g(i) + \frac{c}{2}(\beta - 1)^2 . \qquad (2)$$

By convention, the points derived from the parity data are known as *P*-points and those derived from the fertility rates are known as *F*-points. The goal of the model-fitting procedure is to find a combination of *P*- and *F*-points that are internally consistent with each other (i.e. the two sets of points define essentially the same lines) and then to use these to determine jointly the parameters $\alpha$ and $\beta$ in Equations 1 and 2 above. The values of $\alpha$ and $\beta$ are used to derive the relational gompits, $Y(x) = \alpha + \beta Y^s(x)$, and similarly for $Y(i)$.

Deriving a fitted fertility distribution using the relational Gompertz method requires tabulations of calculated average parities and fertility rates by age. The fertility rates are cumulated and ratios of successive cumulated values are computed. Ratios of successive average parities are also calculated. Gompits of these ratios are calculated and used to plot the two pairs of points, $z(x)$–$e(x)$ against $g(x)$, and $z(i)$–$e(i)$ against $g(i)$. The fitted lines will have slopes equal to $\beta$, and an intercept term involving $\alpha$, $\beta$ and $c$, from which $\alpha$ can be calculated. The values of $\alpha$ and $\beta$ are used to transform the gompits of the standard cumulants into fitted gompits, which are then converted to fitted average parities and fertility rates. The level of fertility is set by the most reliable parity points. These are usually those on women aged 20–29 or 20–34 who are both less likely to omit births and likely to report their ages more accurately than older women.

The use of the relational Gompertz model in the calculation of a fitted fertility distribution has a number of advantages over the earlier *P*/*F* ratio method. The model uses a reliable fertility pattern for medium- to high-fertility regimes (the Booth standard). Thus unreliable fertility rates estimated from reports of births in the last year can be replaced by model values which are fitted using the more reliable points. The plot of the two series of points is a powerful guide to the reliability of each point, and can provide insight into data errors as well as identify fertility trends. All reliable points can be used to derive the fitted model distribution. The model also provides a reliable way of interpolating between values to make parity and cumulated fertility data comparable and to convert fertility rates in unconventional age groups to rates that apply to conventional age groups.

## DATA REQUIREMENTS AND ASSUMPTIONS

### Tabulations of data required

- Fertility rates for the 12, 24 or 36 months before the survey, classified by age of mother at survey, or by age at birth of child; or
  - number of women at the census or survey date, by five-year age group; and
  - number of births to women in the 12, 24 or 36 months before the survey, by five-year age group.
- Average parities of women, classified by five-year age group of mother; or
  - number of women, by five-year age group; and
  - total number of children born to women, by five-year age group.

**Important assumptions**

- The standard fertility schedule chosen for use in the fitting procedure appropriately reflects the shape of the fertility distribution in the population.
- Any changes in fertility have been smooth and gradual and have affected all age groups in a broadly similar way.
- Errors in the pre-adjustment fertility rates are proportionately the same among women in the central age groups (20–39), so that the age pattern of fertility described by reported recent births is reasonably accurate.
- The parities reported by younger women (aged 20–29 or 20–34) are accurate.

The method usually allows violations of these assumptions to be detected.

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Before commencing analysis of fertility levels using this method, analysts should investigate the quality of the data at least in respect of the following dimensions:

- age and sex structure of the population;
- reported births in the last year; and
- average parities and the necessity of an el-Badry correction.

## CAVEATS AND WARNINGS

- In applying this method, analysts must take particular care to ascertain and correctly specify the definition used to classify age of mother.
- Where appropriate and necessary, the average parities should be the corrected average parities after application of the el-Badry correction for the misreporting of childless women as parity not stated.
- The method can handle data aggregated over a three-year period. However, caution should be exercised in using the full model (as opposed to using it simply for smoothing) with data for periods of much longer than a year. Ideally, person-years exposed to risk should be calculated more accurately if using a longer period of investigation. In addition, there is the risk of multiple births occurring within an extended period of investigation, and the form of the questions in the census or survey instrument may be inadequate to the task of identifying such cases.
- If sample or design weights have been provided with the data, they must be applied in the manner appropriate to the statistical software used when deriving the tabulations used as inputs.
- The method is contra-indicated where the shape of the fertility distribution being modelled differs markedly from that of the underlying fertility standard. Since the modelled parameters $\alpha$ and $\beta$ define the shape and location of the fertility schedule, Zaba (1981) recommends that the model only be applied where $-0.3 < \alpha < 0.3$ and $0.8 < \beta < 1.25$. An alternative standard should be considered if $\alpha$ and $\beta$ lie outside these ranges.
- Some of the approximations used in obtaining the estimating equations work less well for the youngest and oldest age groups than for those in the middle age range, especially if the reported fertility schedule is radically different from the standard. The points derived from the reports of these women should therefore be treated with extra caution. However, this has little impact on the estimates of Total Fertility.

## APPLICATION OF METHOD

The method is applied in the following stages.

***Step 1: Calculate the reported average parities***

Calculate the average parities, ${}_5P_x$, of women in each age group $[x, x+5)$, for $x = 15, 20\ldots, 45$, if not already done as part of the preliminary investigations, or produced as a consequence of applying the el-Badry correction. The derivation and correction of average parities is described in Chapters 3 and 4.

***Step 2: Determine the classification of the age of mother***

Depending on the data available, the fertility rates may be classified either by age of mother at the survey date, or by age of mother at birth of her child. The former ages are almost always encountered in the analysis of census data, where the mother's age is her age at the census. The latter are more commonly encountered with administrative data derived from vital registration systems. It is crucial that this classification is determined correctly as mis-specification here will bias the estimated rates produced.

The spreadsheet implementation of the model can accommodate data with no shift (i.e. reported according to the age of mother at birth); or – in the case of data classified by age of mother at survey date – with half a year, a year or one and a half year's shift (for periods of investigation of 12, 24 and 36 months, respectively).

***Step 3: Calculate implied age-specific fertility rates***

Age-specific fertility rates are derived by dividing the births reported in the period of investigation (e.g. the year, two years or three years) before the survey date by the number of women in each age group.

***Step 4: Choose the fertility standard to be applied and the model variant to be fitted***

The default fertility standard is that produced by Booth, modified slightly by Zaba (1981). The standard is appropriate to high- and medium-fertility populations and is a normalized cumulated fertility schedule (i.e. with total fertility equal to one). The standard $Y^s(x)$ values are determined by taking the gompits of the schedule and the standard parity values, $Y^s(i)$, are the gompits of the parities associated with the standard fertility schedule. The choice of standard determines the values of $g()$ and $e()$ used in the regression fitting procedures which are derived algebraically from the $Y^s()$.

Two variants of the relational Gompertz model are presented here. The default option is to make the same assumptions about the nature of errors inherent in fertility data as in the Brass *P*/*F* method, namely that reports of recent fertility suffer from reference-period errors and under-reporting that are independent of age, and that reports of lifetime fertility suffer from omission errors that increase with age. In the spreadsheet, this is referred to as the 'Shape *F* – Level *P*' variant.

The second variant involves using the relational Gompertz model to correct for possible distortions in the *shape* of the fertility distribution, while leaving the level unchanged. Clearly, if reference period errors or under-reporting are suspected, this variant will not give a plausible estimate of fertility.

***Step 5: Evaluate the plot of P-points and F-points***

The plots of $z(x)$–$e(x)$ against $g(x)$, and $z(i)$–$e(i)$ against $g(i)$ on the same set of axes are then used as a diagnostic for identifying common errors and trends in the data (see below).

***Step 6: Fit the model by selecting the points to be used***

Initially, all points should be included in the model, the only exception being if the average parities in one age group are higher than the average parities in the next. In this case the gompit will be undefined and the model cannot be fitted using that point. (Such a situation cannot occur in a real cohort, but could arise because of data error or in a synthetic cohort during a time of rapidly changing fertility.)

If the parity and fertility data are internally consistent, the plots of $z()$–$e()$ against $g()$ should result in straight lines. Those *P*-points and *F*-points that cause each plot to deviate from a straight line should be excluded from the model. Ordinary least squares regression is used to fit lines to the *P*-points and *F*-points and to identify, sequentially, those points that do not fit neatly on a straight line. The intention is to seek the largest combination of *P*- and *F*-points that lie (almost) on the same line, and to use these to fit the model.

Points are selected for inclusion or exclusion using the following guidelines:

- A contiguous series of points must be included in the model. Sequentially, only the end-most points can be excluded. (The reason for this is that each point on the graph is the result of calculations involving the ratio of a pair of adjacent data values. If the analysis leads you to conclude that a data value is unreliable as a denominator of one of these ratios, it is not logical to accept it as the numerator of the next ratio.)
- *P*-points should be eliminated in preference to *F*-points. This is because the average parity data are generally more prone to age-specific errors than the fertility data.
- *P*-points which deviate clearly from the straight line based only on the other *P*-points as well as *F*-points which deviate clearly from the straight line based only on the other *F*-points should be eliminated early on in the fitting process.
- *P*- and *F*-points at older ages should be eliminated in preference to those at younger ages since data at these ages are usually the least reliable and show the least consistency between lifetime and recent fertility. The exception to this relates to the data points for women under the age of 20 because small numbers of events, as expected for younger women, frequently make the estimates of average parities or cumulated fertility unreliable.
- Where only a marginally worse fit is achieved with more points, this is to be preferred to a slightly better fit achieved with fewer points. The spreadsheet calculates the root mean squared error (RMSE),

$$RMSE = \sqrt{\frac{\sum\left(\left(z()-e()\right)-\left(\alpha+(\beta-1)^2\frac{c}{2}+\beta g()\right)\right)^2}{n}},$$

from the points used to fit the model. This statistic can assist with determining the optimal number of data points to which to fit if there is uncertainty as to which of two

competing models is better. In this situation, one can choose the model with the lower RMSE.

***Step 7: Assess the fitted parameters***

The values of $\alpha$ and $\beta$ that represent the best-fitting line joining the remaining *P*-points and *F*-points must be checked to confirm that they are not so far from their central values as to suggest that the standard chosen is inappropriate. A good fit is indicated if $-0.3 < \alpha < 0.3$, and if $0.8 < \beta < 1.25$.

If the parameters lie outside this range, one or both of the underlying data series are problematic or the standard is inappropriate. Experimentation with another standard (see below) or changing the selection of points should be done before proceeding further. If the parameters still lie outside the ranges above, the method should be regarded as inappropriate.

***Step 8: Fitted ASFRs and total fertility***

Having estimated the two parameters of the model, they can be applied to the standard values for the parities to obtain fitted values $Y(i) = \alpha + \beta . Y^s(i)$. These are then converted back into measures of the cumulative proportion of fertility achieved by age group *i* using the anti-gompit transformation. The anti-gompits based on the parity distributions indicate the proportion of fertility achieved by that age group. Dividing the observed parity in each age group by these proportions produces a series of estimates of total fertility. Averaging these values across the sub-set of age groups that were used to estimate $\alpha$ and $\beta$ gives the fitted estimate of total fertility, $\hat{T}$.

Applying the same $\alpha$ and $\beta$ to the standard gompits for the ages that divide conventional age groups (i.e. 20, 25…50), applying the anti-gompit transformation, and multiplying by $\hat{T}$ produces a scaled cumulated fertility schedule. Differencing successive estimates of cumulated fertility and dividing by five produces the fitted fertility schedule for conventional age groups (15–19; 20–24 etc.) even if the data were initially classified with a half-year shift.

(If the model has been fitted using only the *F*-points, then $\alpha$ and $\beta$ are defined by the *F*-line only. The smoothed fertility schedule is produced by a series of steps identical to that described above except that the fitted proportions are multiplied by the level of fertility estimated from the recent data themselves, rather than by an estimate based on the parity data.)

## INTERPRETATION AND DIAGNOSTICS

### Typical errors in the data

The points derived from data applicable to women aged less than 20 are often unreliable, as they are typically derived from fairly small numbers of events and prone to a variety of reporting errors, such as an enumerator ascribing an older age to teenage mothers. It is thus common for the lines fitted to the *P*- and *F*-points to agree for women at peak child-bearing ages (20–34), but not at very young or older ages. If the *P* and *F* lines do not converge even in the 20–34 age range, then either errors must be present in one or both data sets, even at these younger ages, or (substantial) recent fertility changes must have occurred.

A plot of all *P*-points and *F*-points provides information on errors present in the data and recent fertility trends. It is useful when interpreting the plots to remember that the $z()–e()$ values (on the *y*-axis) vary with the observed fertility and parity schedules, whereas the $g()$ values (which are based solely on the standard) do not. Likewise, $z()–e()$ changes in the same direction as the underlying ratios.

The most common types of issues highlighted by the diagnostic plot are omission of children in the parity reports of older women, age exaggeration, and an indication of recent declines in fertility.

Zaba (1981) used simulated data based on the Booth standard to explore the effect of data errors and fertility changes on the plots. The results are described below.

***1) Older women omit children in reporting their lifetime fertility***

If older women omit children in reporting their completed parities, then the *P*-values will tend to be too high (as the denominator of each cumulant will be disproportionately low) relative to the straight-line pattern anticipated and the *P*-points tend to curve upward at older ages.

***2) Exaggeration of births, or age exaggeration by older women***

Both these errors have the same effect, either because an erroneous number of births are reported to older women, or because younger women (who tend to have higher recent fertility) are mistakenly classified as being older than they are in reality. As a result, the *F*-line curves downward at the oldest ages.

### *3) Trends in fertility*

Trends in fertility level are shown by the divergence of *P*- and *F*-points on the graph. If fertility has been falling, the *F*-cumulants tend to be higher than the *P*-cumulants at the same age, and the *F*-points have a steeper slope than the *P*-points. The diagnostic for falling fertility is therefore that the *F*-points tend to lie on a line above that for the *P*-points, and vice versa.

Rapid changes in fertility that have affected the younger ages of childbearing usually prevent the *P*- and *F*-points falling on a common line even when almost all the *P*-points are excluded from the fit. Successive elimination of *P*-points that fails to align the *P*- and *F*-points suggests that fertility has changed rapidly and recently in the younger age groups.

Typical diagnostic plots, based on the Booth standard, are shown in Figure 7.1.

As can be seen, if older women omit live births, the *P*-points on the right hand of the scale (those for older ages) will drift upwards. When women report themselves (or are reported) to be older than they are, the effect is for the *F*-curve to curve downwards at older ages. Finally, if fertility is falling, the *F*-points will generally lie above the *P*-points.

In dealing with real data, one is often faced with a mixture of errors and trends which may be considerably more complicated than the neat archetypes set out here. Severe errors may obscure real trends and, for this reason, the method should not be applied indiscriminately.

### ***P/F* Ratios**

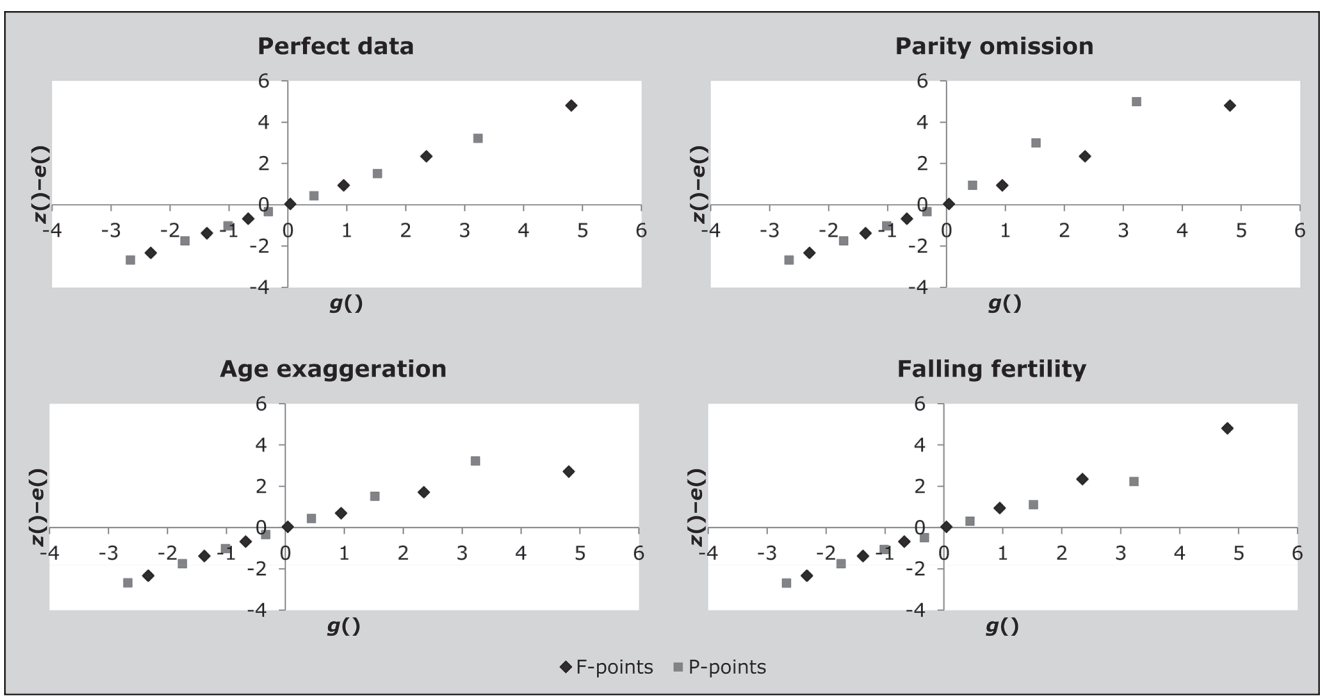
While the *P*/*F* ratio method is not described in this manual, the ratios that result from the application of the method provide useful insights into recent trends in fertility. They can also be used as guides to the applicability of certain methods used to estimate child mortality.

Pseudo *P*/*F* ratios for each age group can be derived quite easily from a fitted relational Gompertz model. The ratio is calculated as

$$\frac{P}{F}(x) = \frac{{}_5P_x}{\hat{F}\cdot\exp\left(-\exp\left(\alpha_F + \beta_F Y^s(x+2.5)\right)\right)},$$
$$x = 20, 25\ldots, 45.$$

The numerator is the observed average parity in each age group while the denominator uses the values of $\alpha$ and $\beta$

*FIGURE 7.1 Diagnostic plots based on simulated data showing common errors*



derived from the *F*-points only (much as in the *F*-only variant of the model) to modify the standard gompit at the mid-point of each age group. The anti-gompit is then scaled up by the level of total fertility implied by the *F*-points selected to be used in the model. The ratio is not calculated for the youngest age group because – typically – average parities are very low as is cumulated fertility to age 17½, thereby causing the ratio to be unstable in that age group.

These *P*/*F* ratios can be plotted in reverse order so that the oldest age group is on the left. The series of *P*/*F* ratios can then be read as running through calendar time from left to right (since, in general, older women's fertility will have occurred earlier in time than younger women's fertility). Excessive deviations from the general trend suggest errors in the data. A downward trend in the *P*/*F* ratios (as plotted) shows increasing divergence between cohort and period fertility measures with increasing age, and hence is indicative of declining fertility.

## WORKED EXAMPLE

This illustration of the method uses data presented in the report on fertility from the 2008 Malawi Census. The method has been implemented in an Excel workbook (see website).

### *Step 1: Calculate the reported average parities*

The average parities are presented in Table 2.6 of the 2008 Malawi Census fertility report. It is not clear from the report whether the parities were edited or whether an el-Badry correction was applied to the data, shown in Table 7.1.

*TABLE 7.1 Measures of fertility from the Malawi 2008 Census*

| Age (at survey) | Average parity per woman | Period fertility rates |
|---|---|---|
| 15–19 | 0.283 | 0.111 |
| 20–24 | 1.532 | 0.245 |
| 25–29 | 2.849 | 0.230 |
| 30–34 | 4.185 | 0.195 |
| 35–39 | 5.214 | 0.147 |
| 40–44 | 6.034 | 0.072 |
| 45–49 | 6.453 | 0.032 |

### *Step 2: Determine the classification of the age of mother*

The question on recent fertility in the 2008 Malawi Census was 'how many live births in the last 12 months'. Since there is no way of dating the child's birth, one can assume that the data are classified by age of mother at the census date rather than at the birth of her child.

### *Step 3: Calculate implied age-specific fertility rates*

Fertility rates are presented in Table 2.6 of the 2008 Malawi Census fertility report. (The derivation of these rates, presented in Table 2.3 of the report, suggests that ${}_5f_{20}$ was 0.250, but the rates in Table 2.6 are retained for the purpose of this example so as to allow a better comparison of the results derived.)

### *Step 4: Choose the standard to be applied and the model variant to be fitted*

In the absence of an alternative, we apply the Booth standard, and – in order to correct the shape *and* level of the fertility data, elect to fit the Shape-*F* Level-*P* variant. The coefficients, $e()$ and $g()$, are derived in Tables 7.2–7.4 below.

Starting with the values from the standard in column [2], the gompits of the standard are calculated in column [3]. For example, in the age group ending at 19½, it is $-\ln(-\ln(0.1140)) = -0.7753$. Note that the cumulated values apply to ages 14½, 19½ etc., reflecting the half-year shift in the classification of mothers' ages. The ratios of successive pairs of cumulated fertility from the standard in column [2] are presented in column [4], and the gompits of these are shown in column [5]. Thus, in the age group ending at age 39½, it is $2.6209 = -\ln(-\ln(0.9298)) = -\ln(-\ln(0.9199/0.9893))$.

The first and second derivatives at the point where $\beta = 1$, presented in columns [6] and [7], are evaluated using the formulae:

$$\frac{d}{d\beta}\phi_x(1) = \phi_x'(1) = \frac{Y^s(x+5).\exp\left(Y^s(x)\right) + Y^s(x).\exp\left(Y^s(x+5)\right)}{\exp(Y^s(x)) - \exp(Y^s(x+5))};$$

$$\frac{d^2}{d\beta^2}\phi_x(1) = \phi_x''(1) = \frac{\left(Y^s(x) - Y^s(x+5)\right)^2.\exp\left(Y^s(x) + Y^s(x+5)\right)}{\left(\exp(Y^s(x)) - \exp(Y^s(x+5))\right)^2}.$$

Finally, $e(x)$ is derived in column [8] by differencing columns [5] and [6]. The $g(x)$ values in column [9] are equal to the values in column [6]. Table 7.3 repeats the calculations, but for unshifted data; these values are required to produce the final, unshifted, fertility estimates. Table 7.4 shows the derivation

*TABLE 7.2 Derivation of e(x) and g(x) when data are subject to a half-year shift*

| Age $x$ [1] | $F^s(x)/F$ [2] | $Y^s(x)$ [3] | Ratio [4] | Phi [5] | Phi' [6] | Phi" [7] | $e(x)$ [8] | $g(x)$ [9] |
|---|---|---|---|---|---|---|---|---|
| | | =gompit[2] | $=Y^s(x)/Y^s(x+5)$ | =gompit[4] | | | =[5]–[6] | =[6] |
| 14 ½ | 0.0011 | –1.9228 | 0.0094 | –1.5410 | –2.4565 | | 0.9155 | –2.4565 |
| 19 ½ | 0.1140 | –0.7753 | 0.3233 | –0.1216 | –1.4527 | 0.9563 | 1.3311 | –1.4527 |
| 24 ½ | 0.3528 | –0.0411 | 0.6007 | 0.6741 | –0.7426 | 0.9632 | 1.4167 | –0.7426 |
| 29 ½ | 0.5872 | 0.6305 | 0.7529 | 1.2592 | –0.0364 | 0.9530 | 1.2957 | –0.0364 |
| 34 ½ | 0.7800 | 1.3925 | 0.8479 | 1.8021 | 0.8405 | | 0.9615 | 0.8405 |
| 39 ½ | 0.9199 | 2.4830 | 0.9298 | 2.6209 | 2.1799 | | 0.4409 | 2.1799 |
| 44 ½ | 0.9893 | 4.5323 | 0.9893 | 4.5324 | 4.5315 | | 0.0010 | 4.5315 |
| Phi"–bar | | | | | | 0.9575 | | |

*TABLE 7.3 Derivation of e(x) and g(x) when data are not subject to age shifting*

| Age $x$ [1] | $F^s(x)/F$ [2] | $Y^s(x)$ [3] | Ratio [4] | Phi [5] | Phi' [6] | Phi" [7] | $e(x)$ [8] | $g(x)$ [9] |
|---|---|---|---|---|---|---|---|---|
| 15 | 0.0028 | –1.7731 | 0.0204 | –1.3591 | –2.3278 | | 0.9688 | –2.3278 |
| 20 | 0.1358 | –0.6913 | 0.3600 | –0.0214 | –1.3753 | 0.9582 | 1.3539 | –1.3753 |
| 25 | 0.3773 | 0.0256 | 0.6200 | 0.7379 | –0.6748 | 0.9629 | 1.4127 | –0.6748 |
| 30 | 0.6086 | 0.7000 | 0.7644 | 1.3143 | 0.0393 | 0.9510 | 1.2750 | 0.0393 |
| 35 | 0.7962 | 1.4787 | 0.8559 | 1.8607 | 0.9450 | | 0.9157 | 0.9450 |
| 40 | 0.9302 | 2.6260 | 0.9378 | 2.7455 | 2.3489 | | 0.3966 | 2.3489 |
| 45 | 0.9919 | 4.8097 | 0.9919 | 4.8098 | 4.8086 | | 0.0012 | 4.8086 |
| Phi"–bar | | | | | | 0.9575 | | |

*TABLE 7.4 Derivation of e(i) and g(i) from parity data*

| Age $i$ [1] | $P^s(i)$ [2] | $Y^s(i)$ [3] | Ratio [4] | Phi [5] | Phi' [6] | Phi" [7] | $e(i)$ [8] | $g(i)$ [9] |
|---|---|---|---|---|---|---|---|---|
| | | =gompit[2] | $=Y^s(i)/Y^s(i+1)$ | =gompit[4] | | | =[5]–[4] | =[6] |
| 0 | 0.0003 | –2.0961 | 0.0056 | –1.6449 | –2.6738 | | 1.0289 | –2.6738 |
| 1 | 0.0521 | –1.0833 | 0.2044 | –0.4622 | –1.7469 | 0.9519 | 1.2846 | –1.7469 |
| 2 | 0.2549 | –0.3124 | 0.5143 | 0.4081 | –1.0159 | 0.9638 | 1.4240 | –1.0159 |
| 3 | 0.4957 | 0.3541 | 0.7014 | 1.0367 | –0.3349 | 0.9597 | 1.3717 | –0.3349 |
| 4 | 0.7067 | 1.0579 | 0.8140 | 1.5810 | 0.4406 | | 1.1404 | 0.4406 |
| 5 | 0.8681 | 1.9561 | 0.8969 | 2.2184 | 1.5162 | | 0.7022 | 1.5162 |
| 6 | 0.9679 | 3.4225 | 0.9701 | 3.4943 | 3.2238 | | 0.2705 | 3.2238 |
| Phi"–bar | | | | | | 0.9585 | | |

of the tabulated values of $e(i)$, $g(i)$ and $c$ for use with the parity data, using the parities from the standard as inputs in column [2].

***Step 5: Evaluate the plot of P-points and F-points***

Following the guidelines above, we begin by fitting models using all the *P*- and *F*-points respectively.

The results are shown in the first plot on the *Diagnostic plots* sheet of the Excel workbook (see website) (Figure 7.2).

While the lines fitted to the *P*-points and the *F*-points lie almost on top of each other, neither fits their underlying data series particularly well. The *F*-points curve downward markedly at the oldest ages, suggesting some degree of age-exaggeration in the data, while the fact that the *P*-points lie just below the *F*-points is an indication that a slight decline in fertility is underway.

***Step 6: Fit the model by selecting the points to be used***

Examination of the plot suggests that a better fit to both lines might be achieved if the *P*- and *F*-points for the last age group were omitted. These points are eliminated from the plot and the resulting revised plot is re-examined (Figure 7.3).

While the lines no longer lie as close together and do not remain parallel, visual inspection suggests that the removal of the next oldest *P*-point might cause all the remaining points to lie on a single line (Figure 7.4).

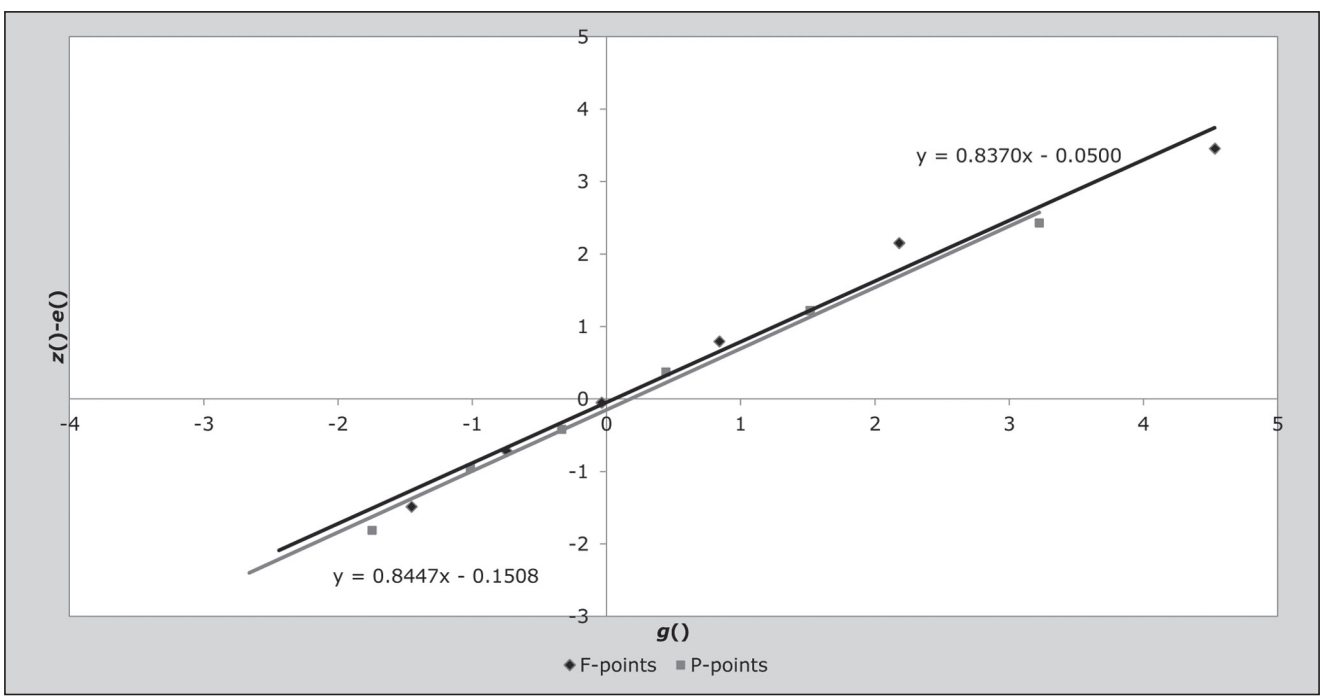
To all intents and purposes, these points can be regarding as falling on a single line, implying that the average parities and fertility rates underlying these points are consistent with each other. No evidence remains that fertility has declined. While a marginally better fit might perhaps be obtained by eliminating the *P*-point associated with the 35–39 age group, to further reduce the number of points included in the model in order to produce a very small improvement in the fit is not worthwhile. Indeed, the exclusion of that *P*-point results in a small increase in the RMSE, from 0.044 to 0.045.

We can accept this fitting of the relational Gompertz model. The third figure in the spreadsheet indicates that the equation of the straight line best fitting the remaining nine data points is $z()–e() = 0.9936.g() – 0.0272$.

From this, the value of $\beta$ is determined directly to be 0.9936, and the value of $\alpha$ is derived from the formula

$$\alpha = -0.0272 - \frac{1}{2}\left(\left(\beta - 1\right)^2 c\right) = -0.0272,$$

*FIGURE 7.2 Plot of z()–e() against g(), all data points, Malawi, 2008 Census*



where $c$ is the average of $\phi_x''(1)$ from Table 7.2 for the current fertility data (since the data are subject initially to a half-year shift), and Table 7.4 for the parity data.

***Step 7: Assess the fitted parameters***

The estimated values of $\alpha$ (–0.0272) and $\beta$ (0.9936) are comfortably close to the standard values of 0 and 1. In aggregate, the slightly negative $\alpha$ shows that the observed fertility distribution for Malawi in 2008 is slightly older than the standard and the value of $\beta$ less than 1 suggests that the spread of the distribution is slightly wider than that of the standard.

***Step 8: Fitted ASFRs and total fertility***

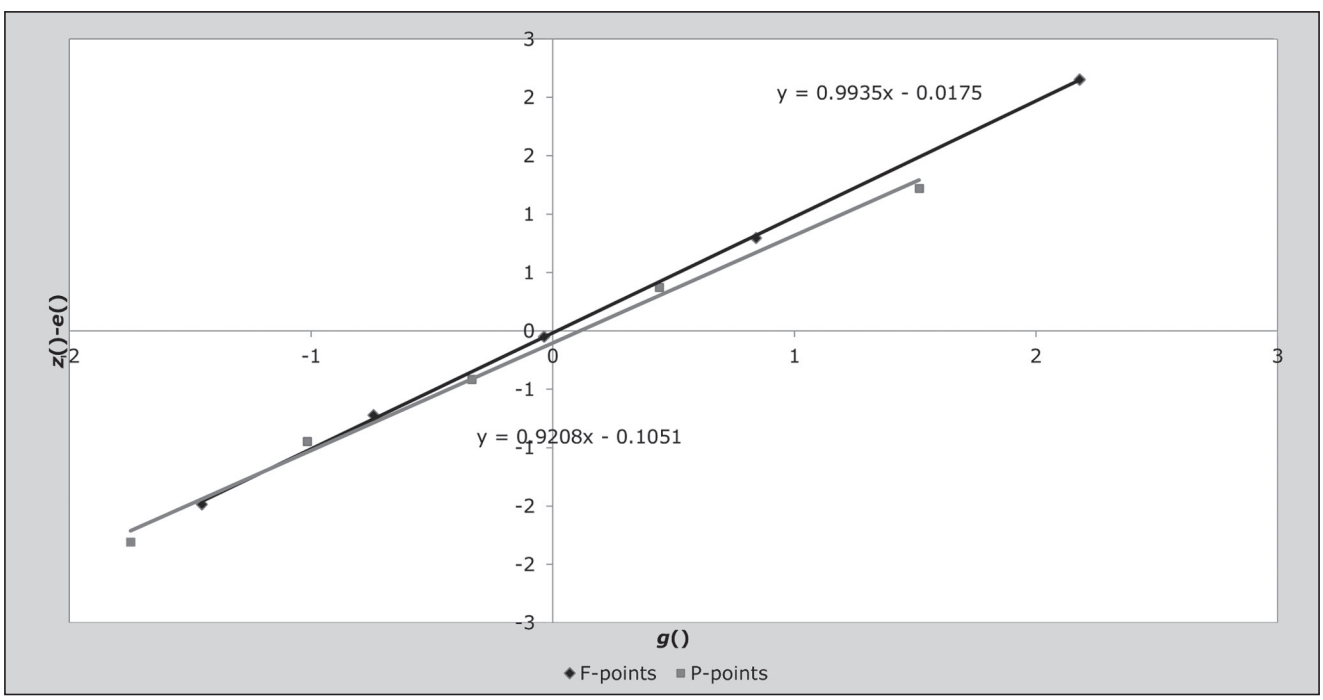
To determine the overall level of fertility, the fitted values of $\alpha$ and $\beta$ are applied to the standard parity gompits (column [3] of Table 7.4) for the age groups *whose P-points were included in the model*, and the anti-gompits calculated (Table 7.5). Dividing the observed average parities for a given age group by the fitted anti-gompit gives the level of fertility implied by the average parities (column [6] of Table 7.5, and an estimate of total fertility, $\hat{T}$, is derived from the arithmetical average of these estimates (= 5.9784).

*TABLE 7.5 Calculation of estimated total fertility, $\hat{T}$, Malawi, 1998 Census*

| Age group $i$ [1] | $Y^s(i)$ [2] | $Y(i)$ [3] | Anti-gompit [4] | $P(i)$ [5] | Implied fertility level [6] |
|---|---|---|---|---|---|
| | | $=\alpha+\beta Y^s(i)$ | | | |
| 1 | –1.0833 | –1.1034 | 0.0491 | 0.283 | 5.7662 |
| 2 | –0.3124 | –0.3375 | 0.2462 | 1.532 | 6.2218 |
| 3 | 0.3541 | 0.3246 | 0.4854 | 2.849 | 5.8694 |
| 4 | 1.0579 | 1.0239 | 0.6982 | 4.185 | 5.9937 |
| 5 | 1.9561 | 1.9162 | 0.8631 | 5.214 | 6.0407 |
| T-hat | | | | | 5.9784 |

To get the associated age-specific fertility rates in conventional ages, we again apply $\alpha$ and $\beta$, but this time to the current fertility gompits, $Y^s(x)$, in Table 7.3. Taking the anti-gompit of the fitted values produces a cumulative fertility distribution. These proportions are multiplied up by the estimate of $\hat{T}$ from the previous step to produce the absolute cumulated fertility distribution. Differencing and dividing by 5 produces the final age-specific fertility rates (Table 7.6).

*FIGURE 7.3 Plot of z()–e() against g() with F- and P-data points associated with the 45–49 age group removed, Malawi, 2008 Census*



The ASFRs are shown in the last column of Table 7.6, with an implied total fertility (15–49) of 5.96 children per woman.

## DETAILED DESCRIPTION OF METHOD

### Introduction

The relational Gompertz model evolved from the Brass *P/F* ratio method. It works with the same input data, and makes use of the parity data from younger women to set the level of fertility, while the shape of the fertility distribution is determined by women's reports on recent births.

### Mathematical exposition

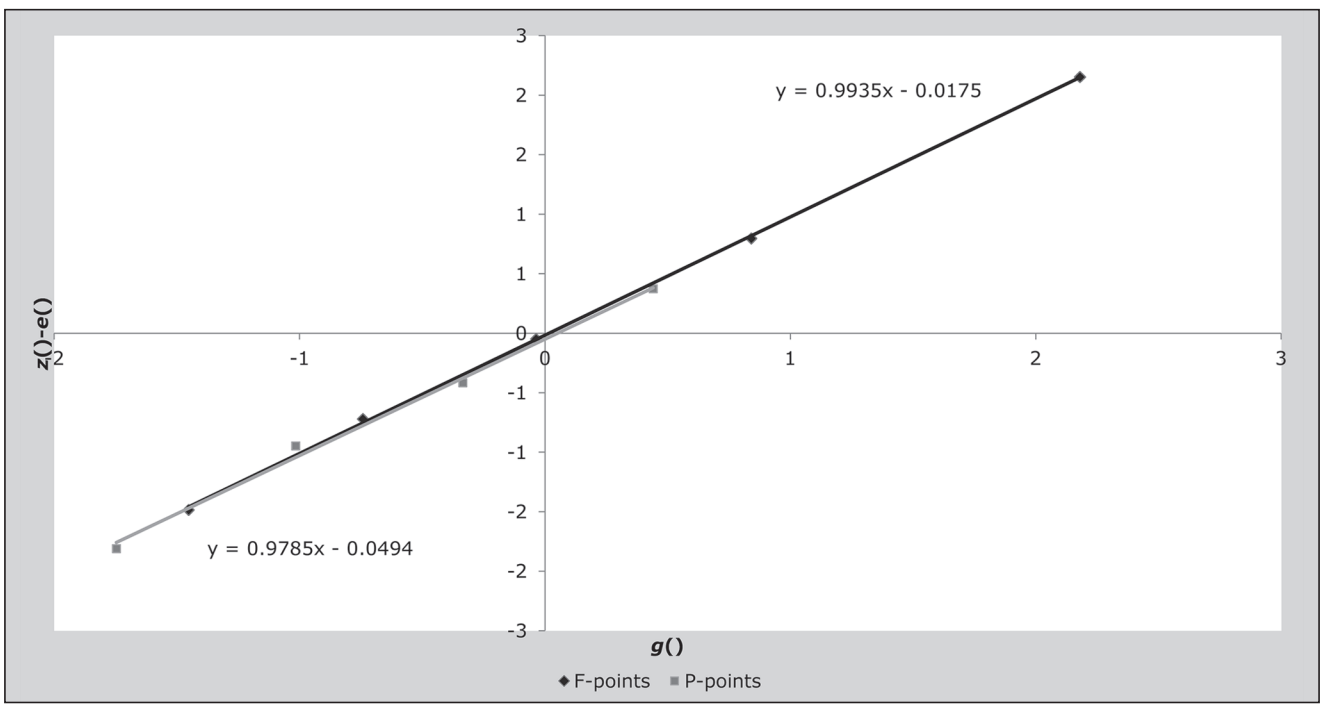
The relational Gompertz model of fertility, initially developed by Brass (1978), is analogous in many ways to the logit models of mortality. The model can be used to describe any fertility distribution by reference to a standard fertility distribution and the parameters used to transform it to produce the required distribution. The transformation used as a basis for the relationship between the two fertility distributions is known as the Gompertz transformation. In the original formulation of the model, it is performed on a cumulated, proportional distribution i.e. where each parity or fertility cumulant is expressed as a proportion of the total fertility rate (the summed distribution). The summed distribution therefore takes a value of one. The transformed proportions are known as gompits and are given by

$$Y(x) = \text{gompit}\left(\frac{F(x)}{F}\right) = -\ln\left(-\ln\left(\frac{F(x)}{F}\right)\right),$$

where $F(x)$ is the sum of the age-specific fertility rates cumulated to age $x$ and $F$ is the total fertility rate. Exactly the same relationship holds for parities, replacing $F(x)$ with average parity for age groups, and $F$ with the cumulated parity at age 50+. The Gompertz transformation 'stretches' the original age axis so that the gompits plotted against age almost form a straight line. However, the transformation is not perfect; the line tends to curve slightly at both ends, as can be seen in Figure 7.5, which plots the fertility rates from Booth's (1984) standard.

The transformation can be used as a basis for a relational model because plots of the gompits of different sets of fertility rates against age tend to deviate from linearity in similar ways and, therefore, the relationship between the two such

*FIGURE 7.4 Plot of z()–e() against g() with P-data point associated with the 40–44 age group removed, Malawi 2008 Census*



sets of gompits themselves is usually close to linear. Using the model in a relational form enables the model parameters to be estimated by fitting straight lines, which is a straightforward process and makes it simpler to interpret the results.

As the gompits of the fertility cumulants of any two fertility distributions have an approximately linear relationship, one can relate the gompits of an observed fertility distribution to the gompits of a standard distribution based on accurate data, by means of the relation

$$Y(x) = \alpha + \beta Y^s(x),$$

where $Y(x)$ is the gompit of cumulated proportionate fertility at age $x$, and $Y^s(x)$ is the gompit of the standard fertility cumulants.

In this formulation, $\alpha$ represents the extent to which childbearing ages in the population differ from the standard with negative values of $\alpha$ making the age schedule of fertility older. $\beta$ represents the extent to which the spread of childbearing differs from the spread in the standard population. The spread of the distribution is narrower for values greater than 1.

The model is, in fact, a three-parameter model. Converting the fitted gompits back to estimates of cumulative fertility using the reverse transformation produces a proportional distribution which sums to one. A third parameter is required to multiply all the fitted values up to the appropriate level of fertility. This is effectively Total Fertility – the very thing one is trying to estimate – but the estimate based on the observed data may not be reliable due to reporting errors. Thus the original fitting procedure (not described here) was adapted by Zaba (1981), whose contributions and extensions to the method are described below.

Zaba's approach uses the gompits of the ratio of adjacent cumulants of fertility to isolate the estimation of the shape parameters from estimation of the level of fertility:

$$Y(x) = \text{gompit}\left(\frac{F(x)}{F(x+5)}\right) = -\ln\left(-\ln\left(\frac{F(x)}{F(x+5)}\right)\right).$$

If the cumulant, $F(x)$, conforms to a Gompertz model with parameters $\alpha$ and $\beta$, then

$$\begin{aligned} & Y(x) \\ & = -\ln\left(\exp\left(-\left(\alpha + \beta Y^s(x)\right)\right) - \exp\left(-\left(\alpha + \beta Y^s(x+5)\right)\right)\right) \\ & = \alpha - \ln\left(\exp\left(-\beta Y^s(x)\right) - \exp\left(-\beta Y^s(x+5)\right)\right) \\ & = \alpha + \phi_x(\beta) \end{aligned}$$

where $\phi_x(\beta)$ is the second term in the penultimate line. For values of $\beta$ close to 1, $\phi_x(\beta)$ can be approximated by a Taylor series expansion about $\beta = 1$:

$$\phi_x(\beta) = \phi_x(1) + (\beta - 1)\frac{d}{d\beta}\phi_x(1) + \frac{(\beta-1)^2}{2}\frac{d^2}{d\beta^2}\phi_x(1) + \ldots \quad (3)$$

From the definition of $\phi_x(\beta)$, $\phi_x(1) = Y(x)$. Further, it can be shown that

*TABLE 7.6 Calculation of corrected fertility rates, Malawi, 1998 Census*

| Age group (up to age $x$) [1] | $Y^s(i)$ [2] | $Y(i)$ [3] | Anti-gompit [4] | Scaled by T-hat [5] | ASFR [6] |
|---|---|---|---|---|---|
| | | $=\alpha+\beta\ Y^s(i)$ | | $F(x)=$ [4]*5.9784 | ${}_5f_{x-5} = (F(x)-F(x-5))/5$ |
| 15 | –1.7731 | –1.7887 | 0.0025 | 0.0151 | 0.0030 |
| 20 | –0.6913 | –0.7140 | 0.1298 | 0.7758 | 0.1521 |
| 25 | 0.0256 | –0.0017 | 0.3673 | 2.1956 | 0.2840 |
| 30 | 0.7000 | 0.6683 | 0.5989 | 3.5807 | 0.2770 |
| 35 | 1.4787 | 1.4419 | 0.7894 | 4.7194 | 0.2277 |
| 40 | 2.6260 | 2.5817 | 0.9271 | 5.5428 | 0.1647 |
| 45 | 4.8097 | 4.7512 | 0.9914 | 5.9269 | 0.0768 |
| 50 | 13.8155 | 13.6984 | 1.0000 | 5.9784 | 0.0103 |

$$\frac{d}{d\beta}\phi_x(1) = \phi_x'(1)$$

$$= \frac{Y^s(x+5).\exp\left(Y^s(x)\right) + Y^s(x).\exp\left(Y^s(x+5)\right)}{\exp(Y^s(x)) - \exp(Y^s(x+5))};$$

$$\frac{d^2}{d\beta^2}\phi_x(1) = \phi_x''(1)$$

$$= \frac{\left(Y^s(x) - Y^s(x+5)\right)^2.\exp\left(Y^s(x) + Y^s(x+5)\right)}{\left(\exp(Y^s(x)) - \exp(Y^s(x+5))\right)^2} \quad (4)$$

Zaba (1981) evaluated this last quantity for a variety of different values of *x*, and showed that it is almost constant in the range $15 \le x < 30$. (This can also be seen in Tables 7.2–7.4 where this quantity is derived). Thus, one can replace $\phi_x''(1)$ by *c*, the arithmetical mean of the quantities in that age range and rewrite Equation 3 as

$$Y(x) = \alpha + \phi_x(1) + (\beta - 1)\phi_x'(1) + (\beta - 1)^2 \frac{c}{2}$$

or as

$$Y(x) + \phi_x'(1) - \phi_x(1) = \alpha + (\beta - 1)^2 \frac{c}{2} + \beta\phi_x'(1).$$

In other words, there is a linear relationship between $Y(x) + \phi_x'(1) - \phi_x(1)$ and $\phi_x'(1)$.

In subsequent work (Sloggett, Brass, Eldridge et al. 1994), Zaba re-expressed these terms as follows:

| Term | Redefined term |
|---|---|
| $\phi_x'(1)$ | $g(x)$ |
| $\phi_x(1) - \phi_x'(1)$ | $e(x)$ |
| $Y(x)$ | $z(x)$ |

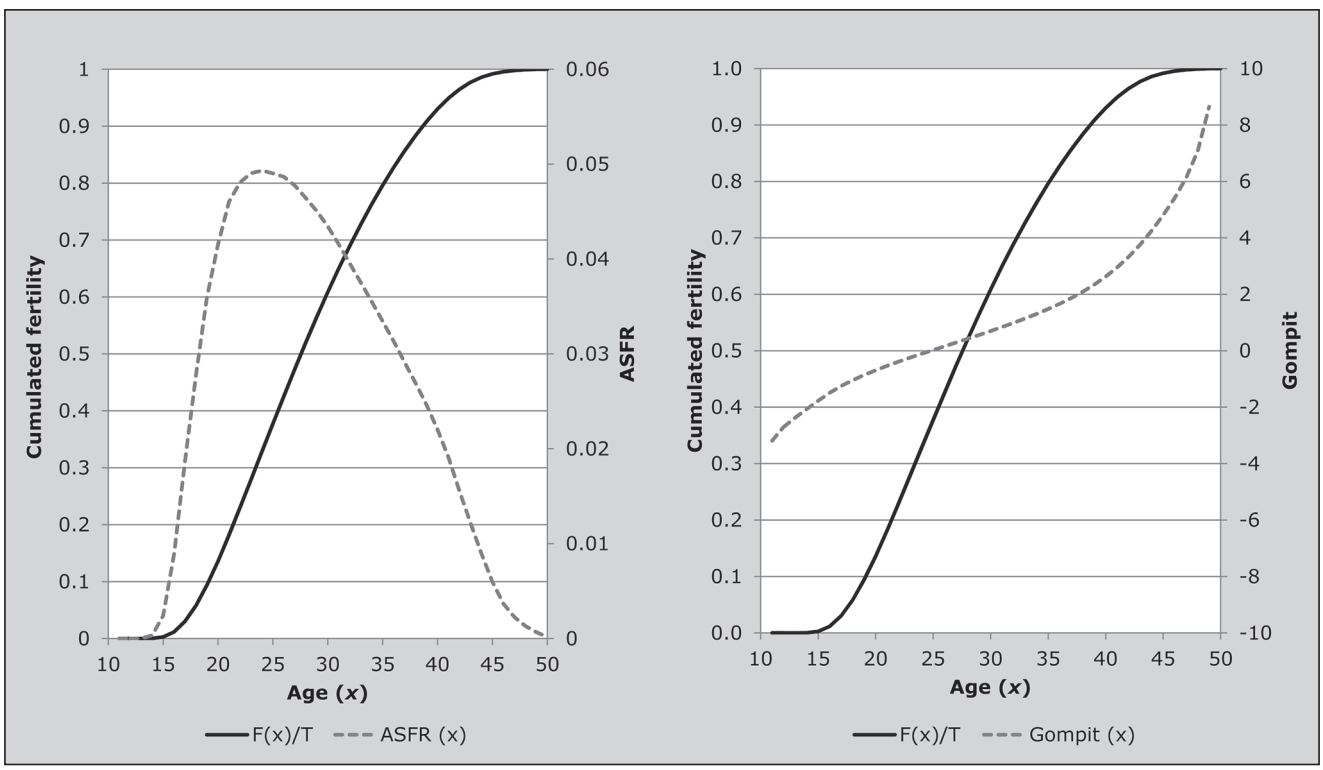
Hence, in this revised notation,

$$z(x) - e(x) = \alpha + (\beta - 1)^2 \frac{c}{2} + \beta g(x)\ ,$$

implying a linear relationship between *z*–*e* and *g*.

Applying the same reasoning as above, the equivalent

*FIGURE 7.5 Effect of the gompit transform on a fertility distribution*



formulation can be derived for $z(i)$–$e(i)$ in terms of $\alpha$, $\beta$, $c$ and $g(i)$.

**Variants of the fitting procedure**

While the standard version of the model set out here uses data on recent fertility to determine the shape of the fertility schedule and sets the level by reference to the (selected) parity points, other variants are possible that privilege one set of input data over the other in different ways. The one presented here uses only the data on recent fertility.

The *F-only* variant privileges the data on recent fertility, and uses them to set both the shape and level of fertility in the model. This variant should, therefore, only be used if the analyst lacks parity data or does not wish them to influence the fit of the model. Thus, this variant simply smoothes the observed fertility rates using a relational Gompertz model.

Another extension of the relational Gompertz model that uses only the data on parity is used to estimate fertility from cohort parity increments. There is also a modified version of the relational Gompertz model making use of data from two censuses or surveys, that produces an estimate of intersurvey fertility from these data.

## CONSTRUCTION OF STANDARDS

**The Booth Standard**

The derivation of the Booth standard is described in detail in Booth (1984). The important aspects associated with the standard and its use in the relational Gompertz model are, first, that the standard is intended for use in medium- to high-fertility populations. Second, the standard was derived from a number of schedules produced by the Coale-Trussell fertility model, and is thus subject to the constraints imposed by that model. For the most part, these are not material.

The standard used here is not identical to that published by Booth. First, Zaba's (1981) standard differs slightly from Booth's below age 15 to obtain a better fit for very early patterns of childbearing. Accommodating these, it is possible to reconstruct fully the tabulated coefficients presented in Zaba (1981) and Sloggett, Brass, Eldridge *et al.* (1994). The standard used here is identical for the unshifted coefficients. Where the shift is required, small differences emerge, arising from the manner in which the original Booth standard has been interpolated. Zaba (1981) calculated the values for $F(x+1/2)$, $F(x+3/2)$, etc. by interpolating between successive values of $F(x)$, $F(x+1)$, $F(x+2)$. However, as the gompit transform linearizes $F(x)$, it makes more sense to interpolate the gompits of $F(x)$, $Y(x)$ for half-year ages and then to establish the values of $F(x+1/2)$, $F(x+3/2)$ etc. by taking the appropriate anti-gompits.

**Construction of alternative standards**

As already noted, the Booth standard was designed for use in medium-high fertility countries. In applications of the relational Gompertz model to low-fertility countries or those with very different patterns of fertility, alternative standards are called for. We describe here briefly how to derive alternative standards.

The basic approach to constructing any standard requires a set of $F(x)$ which can be converted by means of a gompit into a series of $Y(x)$, and then to derive values of $\phi_x(1)$, $\phi_x'(1)$ and $\phi_x''(1)$ from them using the relationships established in Equations 3 and 4. From these, tabulations of $z()$, $e()$ and $g()$ can be calculated. As described above, the values of $\phi_x''(1)$ are almost constant between 15 and 30 for a given standard, and so the three values (15–19; 20–24; 25–29) are averaged to produce estimates of the constant term, $c$.

To construct a new standard, one should begin with an accurate series of age-specific fertility rates, $f^s(x)$. Using conventional demographic analysis, we can then define the equivalent cumulants as

$$F^s(x) = \int_0^a f^s(a)da\,.$$

In most situations $f(a)$ is not an integrable function, so numerical techniques have to be used to approximate the integral closely. Recursively, using the composite trapezium rule,

$$F^s(x) \approx F^s(x-1) + \frac{1}{2}\left(f^s(x-1) + f^s(x)\right).$$

From this, the gompit, $z(x)$ is readily calculated, $z(x) = -\ln(-\ln(F(x)))$.

Using the properties of a Taylor expansion described in Equation 4, the components of $e(x)$ and $g(x)$ can be defined and expressions for these quantities derived.

The values of $z(i)$, $g(i)$ and $e(i)$ are defined similarly, with the only extension being the requirement to derive the constant fertility parities associated with $F(x)$. The parities in any given age group $[x, x+n)$ are given by

$${}_nP_x = \int_x^{x+n} F(a)da$$

which can also be evaluated using a composite trapezium rule.

## FURTHER READING AND REFERENCES

Other than the source material referred to already, literature on the relational Gompertz model is sparse. While this is no doubt due in part to its being described (Booth 1984) shortly after the appearance of *Manual X*, a coherent exposition of how to apply the model appeared only in the SIAP manual (Sloggett, Brass, Eldridge *et al.* 1994). The method has been applied in numerous situations around the globe, although not in the form described here.

The PASEX suite of spreadsheets prepared by the US Census Bureau (1997), for example, offers a somewhat simplified version of the model, forcing the user to fit the straight lines to *P* and *F* using either just two *P*-points and two *F*-points, or three of each, with little regard for the internal consistency of the points chosen. This is the route adopted by the Malawian National Statistics Office in their analysis of fertility data from the 2008 Census. Given the high degree of consistency in these data for all women aged less than forty, the results presented in that report (TFR = 6.0) do not differ in any meaningful way from those presented in the worked example. With less-well behaved data, such congruence of results between the applications should not be taken for granted.


Booth H. 1980. "The estimation of fertility from incomplete cohort data by means of the transformed Gompertz model." Unpublished PhD thesis, London: University of London.

Booth H. 1984. "Transforming Gompertz' function for fertility analysis: The development of a standard for the relational Gompertz function", *Population Studies* **38**(3):495–506. doi: 10.2307/2174137

Brass W. 1978. *The relational Gompertz model of fertility by age of woman*. London: Centre for Population Studies, London School of Hygiene and Tropical Medicine.

Sloggett A, W Brass, SM Eldridge, IM Timæus, P Ward and B Zaba (eds). 1994. *Estimation of Demographic Parameters from Census Data*. Tokyo: Statistical Institute for Asia and the Pacific.

US Census Bureau. 1997. *Population Analysis Spreadsheets for Excel*. Washington, DC: US Census Bureau. http://www.census.gov/population/international/software/pas

Zaba B. 1981. *Use of the Relational Gompertz Model in Analysing Fertility Data Collected in Retrospective Surveys*. Centre for Population Studies Research Paper 81–2. London: Centre for Population Studies, London School of Hygiene & Tropical Medicine.

# Chapter 8 Parity progression ratios

*Tom A Moultrie and Basia Zaba*

When women or parents think about childbearing they usually think in terms of starting a family, or increasing their family by adding another child. This chapter focuses on analysing the incremental aspects of childbearing, looking at proportions of mothers who reach a given parity or birth order, and the proportion who then proceed to increase their parity by at least one more child.

Information on birth order is easily obtainable from two basic census questions: on women's completed parity and on births in the past year. Using the first question one can disaggregate the births in the past year by birth order. Examination of fertility data by parity through the use of parity progression ratios and their projected equivalents gives additional information on childbearing trends and can be used to assess changes in the parity distribution of fertility.

This chapter describes the procedure for calculating parity distributions and parity progression ratios for women who have reached the end of childbearing as well as the derivation of projected parity progression ratios to forecast the eventual distribution of younger women by the parities that they are expected to attain when they reach the end of childbearing.

Parity progression ratios (albeit with somewhat different notation) are well described in the demographic literature (see, for example, Preston, Heuveline and Guillot (2001)). The idea of calculating projected parity progression ratios proposed by Brass (1985) is less widely known. The exposition here focuses on the calculation of the latter quantities, although conventional parity progression ratios are required as part of the process and are described as well.

A parity progression ratio (PPR) is the proportion of women who progress from one parity to the next. PPRs can be calculated for cohorts of women defined either by age or marriage. Usually age cohorts of women are considered i.e. the parity progression ratios are calculated from the parity distribution of a particular age group of women. For cohorts of women that have finished childbearing, and assuming there is no differential mortality by parity of older women, these measures are fixed. For cohorts that are still in the childbearing ages, however, the measures change as increasing numbers of women move to higher parities. The measures derived from younger women thus suffer from both censoring and selection effects as women predisposed to having more children faster will be disproportionately represented in age-parity combinations before the end of childbearing. This means that comparison of the PPRs of younger and older women is misleading.

Brass (1985) describes a technique for projecting PPRs to the end of women's reproductive years that enables one to use the parity data on younger women. The advantage of disaggregating fertility by parity is that different PPRs may respond to different factors, so that changes in particular PPRs may provide insight into processes of fertility change going on in the population. Thus, for example, parity-specific fertility limitation (or its absence) is readily measured by parity progression ratios. PPRs may also be less affected than more common fertility metrics by some types of data error (e.g. reference period errors). The calculation of first birth rates, carried out in the construction of projected PPRs, can also provide an indication of data quality and fertility change.

The calculation of PPRs for cohorts that have completed their fertility is straightforward. Comparison of successive cohorts can give information on trends in fertility, although more reliable conclusions can be drawn if PPRs for the same cohorts can be compared across more than one census. For younger women, who provide information on more recent fertility trends, a more elaborate procedure is required. These rates are used to project the expected parity distribution that the younger women would achieve if they experienced the age-order specific fertility rates (AOSFRs) until the end of childbearing. This parity distribution is used to calculate projected PPRs, which can be used to infer changes in

fertility that are implicit in women's childbearing histories to date combined with the current AOSFRs. An advantage of using the projected PPRs is that, since they are ratios, errors in the AOSFRs that result from under-reporting will tend to cancel out when they are used to project PPRs, with the exception of progression from nulliparity to the first birth.

Two new concepts are introduced. The first is $M(i)$ the proportion ever-attaining parity order $i$, which is the proportion of women who have at least $i$ children. The second is $a(i)$, the parity progression ratio, which measures the proportion of women who progress from a given parity, $i$, to the next parity, $i+1$. Projected proportions attaining parity $i$, $M^*(i)$ and projected parity progression ratios $a^*(i)$ respectively denote the proportions of women who are expected to attain at least parity $i$, and who are expected to progress from parity $i$ to parity $i+1$ by the end of their childbearing years.

## ESTIMATES OBTAINED

The following measures can be calculated specifically for any age group of women:

- The observed proportion ever-attaining parity order $i$, ${}_5M_x(i)$, is the proportion of women aged $x$ to $x+5$ who have had $i$ or more births. The corresponding projected proportion ${}_5M_x^*(i)$ is the proportion of women currently aged $x$ to $x+5$ who are expected to have had $i$ or more births by the end of their reproductive lives. After the end of childbearing, for $x \geq 50$, the observed and expected proportions are identical.
- Observed parity progression ratios of order $i$ for women aged $x$ to $x+5$ are denoted ${}_5a_x(i)$ to represent the proportion of women with at least $i$ children who have progressed to have at least one more subsequent birth. Projected parity progression ratios are denoted by ${}_5a_x^*(i)$. Once again, for $x \geq 50$ observed and expected ratios are identical.

The following identities link these measures, and relate them to the distribution of women by parity.

Define $N(i)$ to be the number of women in the population of parity $i$ exactly, and denote the total number of women in the population as $N$.

Let $W(i)$ be the number of women in the population having attained parity $i$ or higher, and denote the highest parity attained in the population by $\pi$.

The number who have attained parity $i$ or higher is obtained by summing higher order $N(i)$:

$$W(i) = \sum_{j=i}^{\pi} N(j) = N(i) + N(i+1) + \ldots + N(\pi)$$

so that the corresponding proportion, $M(i)$, is

$$M(i) = \frac{1}{N} \cdot \sum_{j=i}^{\pi} N(j).$$

The number of women of parity zero and over is

$$W(0) = \sum_{j=0}^{\pi} N(j) = N$$

so that the proportion at parity zero and over, $M(0) = N/N = 1$.

The average parity, $P$, in the whole population is

$$P = \frac{1}{N} \cdot \sum_{j=1}^{\pi} j \cdot N(j) = \frac{N(1)}{N} + \frac{2 \cdot N(2)}{N} + \frac{3 \cdot N(3)}{N} + \ldots + \frac{\pi \cdot N(\pi)}{N} = \frac{1}{N} \cdot \sum_{j=1}^{\pi} W(j) = \sum_{j=1}^{\pi} M(j).$$

For women who have completed childbearing, average parity is the same as completed family size, *CFS*, the equivalent of Total Fertility in a real cohort:

$$CFS(x \text{ to } x+5) = \sum_{j=1}^{\pi} {}_5M_x(j) \text{ for } x \geq 50.$$

The parity progression ratios can be written in terms of numbers or proportions:

$$a(i) = \frac{W(i+1)}{W(i)} = \frac{W(i+1)/N}{W(i)/N} = \frac{M(i+1)}{M(i)}.$$

Conversely, the proportion of women attaining parity $i+1$ or higher can be expressed in terms of parity progression ratios a($i$) for lower birth-order transitions:

$$\begin{aligned} M(i+1) &= a(i) \cdot M(i) = a(i) \cdot a(i-1) \cdot M(i-2) \\ &= a(i) \cdot a(i-1) \cdot a(i-2) \cdot \ldots \cdot a(0) \\ &= \prod_{j=1}^{i} a(j) \end{aligned}$$

where the symbol $\prod$ denotes cumulative multiplication.

## DATA REQUIREMENTS AND ASSUMPTIONS

### Tabulations of data required

- Observed party progression ratios
  - Parity by age group of women aged 45–49 or more.

- Projected parity progression ratios
  - Parity by age group of women aged 45–49 or less; and
  - Number of children born during the year preceding the census, classified by mothers' age (in five-year groups) and number of children ever born. Given the assumptions described below, exact apportionment of births in the last year is not called for. Use of births in each of the 12 *complete* months preceding the census date is recommended.

**Important assumptions**

- Women have had at most one birth in the past year.
- The AOSFRs derived will continue to apply in the future. This assumption may be unrealistic. It indicates that projected PPRs measure the implications of women's achieved fertility in combination with current AOSFRs. The PPRs should not be treated as forecasts for cohorts likely to have a considerable number of additional children of the parity concerned, or where fertility is changing rapidly.

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Before commencing analysis of fertility levels using this method, analysts should investigate the quality of the data at least in respect of the following dimensions:

- age and sex structure of the population;
- reported births in the last year; and
- average parities and the necessity of an el-Badry correction.

## CAVEATS AND WARNINGS

- Where appropriate and necessary, the parities should be those estimated by means of the application of the el-Badry correction for the misreporting of childless women as parity not stated.
- If sample or design weights have been provided with the data, remember to apply them in the manner appropriate to the statistical software used when deriving the tabulations used as inputs.
- Projected PPRs should only be calculated from censuses and other data sets that include a large number of women (at least 10,000 in each age group) because the derivation of reliable age-order-specific fertility rates (which are required to derive the projected parity progression ratio) requires robust estimates of the rates at each age and parity combination.

## APPLICATION OF METHOD

The calculation of projected PPRs involves a number of steps, but the principle behind the calculation is not difficult to grasp. The AOSFRs that are calculated using the births in the previous year are used to project forwards the parity distribution that would be expected for each cohort if those rates were to continue for the rest of the cohort's childbearing. This is done by calculating the order-specific equivalent of total fertility, that is, the cumulated fertility rates for births of the *i*-th order, cumulated across all age groups. These rates are used to calculate the additional births of order *i* expected to be added to each age cohort of women in the interval from their current age to the end of their reproductive span. This is a synthetic cohort measure since it cumulates current period rates. The additional proportion expected to reach parity *n* or more is added to the current proportion of women of birth order *i* or more, to give the final proportion expected to be of parity *i* or more. These cumulated projected parity distributions are used to calculate projected PPRs, which can be examined for fertility trends. Symbolically, and defining ${}_5P_x$ to be the average parity of women aged between $x$ and $x+5$, the projected completed cohort fertility of women aged $x$ to $x+5$ is thus

$$ {}_5PTFR_x = {}_5P_x + (TFR - 5\sum_{i=15,5}^{x} {}_5ASFR_i). $$

### *Step 1: Extract a tabulation of children ever born by age group of mother for each age group for which there are data*

Tabulate the number of women in each age group $[x,x+5)$, ${}_5N_x$, by number of children ever born, $i$, to produce a series ${}_5N_x(i)$. If necessary, the data should have been edited to eliminate implausible parities. An el-Badry correction should have been applied if this is indicated.

### *Step 2: Extract a tabulation showing the number of births in the last year, by mother's age and parity, ${}_5B_x(i)$*

The assumption is made at this stage that women have had a maximum of one child in the last 12 months, and that there have been neither multiple deliveries nor multiple confinements in the period. It follows that an implicit assumption is that a birth in the last year to a woman currently of parity $i$ was a birth of order $i$.

***Step 3: Derive proportions ever attaining parity i and parity progression ratios for each age group***

The number of women aged between $x$ and $x+5$ who have given birth to $i$ or more children is thus given by

$$ {}_5W_x(i) = \sum_{j=i}^{\pi} {}_5N_x(j). $$

The proportions of all women aged $[x, x+5)$ who have had $i$ or more births are calculated as

$$ {}_5M_x(i) = \frac{1}{{}_5N_x} \cdot \sum_{j=i}^{\pi} {}_5N_x(j). $$

Proportions with $i$ or more births should be derived for all cohorts, whether or not they have finished their child-bearing. Women of zero parity should be included in the denominator, but those of unknown parity (after the application of an el-Badry correction, if necessary) should not.

The Parity Progression Ratio (PPR) between parity $i$ and $i+1$ is then defined as

$$ {}_5a_x(i) = \frac{{}_5M_x(i+1)}{{}_5M_x(i)}. $$

Derivation of parity progression ratios for women who have completed childbearing ends with this step. The following steps describe how to derive projected parity progression ratios for younger age cohorts of women.

***Step 4: Derive age-order specific fertility rates***

From the tabulation of births in the last year by age group and parity of mother (in Step 2), calculate age-order specific fertility rates for women in each age group 15–19,… 45–49,

$$ {}_5AOSFR_x(i) = \frac{{}_5B_x(i)}{{}_5N_x}. $$

Again, women of unstated parity should be excluded from the denominator. (Note that since the denominator is the total number of women in the age group (${}_5N_x$), and hence not a measure of exposure to risk of generating the events in the numerator, the AOSFRs are not true incidence rates).

***Step 5: Derive cumulated order-specific fertility rates***

Cumulated age-order specific fertility rates up to age $x+5$ for order $i$ are calculated from

$$ {}_5TOFR_x(i) = 5. \sum_{j=15,5}^{x} {}_5AOSFR_j(i). $$

It follows that ${}_5TOFR_{45}(i) - {}_5TOFR_x(i)$ is a measure of the additional proportion of women expected to achieve parity $i$ between age $x+5$ and the end of the reproductive period, on the assumption that future fertility will remain the same as current fertility.

***Step 6: Calculate the age distribution of order-specific fertility rates, and interpolate to conventional ages***

The cumulated rates derived in the previous step apply to ages 19.5, 24.5, etc. (the half-year shift being on account of the classification of mother's age being age at the census date, and not her age at the birth of the child). The proportions ever attaining each parity and parity progression ratios, on the other hand, apply (roughly) to the mid-point of each age group 17.5, 22.5, etc. The age-order rates must therefore be interpolated so that they apply to the central age of each age group. Interpolation is carried out by expressing the cumulated age-order fertility rates as proportions of the total order-specific rates and using linear interpolation on the gompits of the cumulants.

For each birth order, calculate the proportion of the total order-specific rate achieved by the upper limit of the age group. Since rates for shifted age groups have been used in this example, the upper limits are shifted down by half a year. The proportions are calculated by summing the rates up to the upper limit of the age group, and expressing this summation as a proportion of the total rate i.e.

$$ {}_5\theta_x(i) = \frac{{}_5TOFR_x(i)}{{}_5TOFR_{45}(i)}. $$

As already noted, these proportions apply to ages 19½, 24½, etc. Due to the curvature of the fertility schedule, direct linear interpolation to obtain the proportions at ages 22½, 27½ etc. is not appropriate. A gompit (double negative log) transform, however, transforms a sigmoidal function to a straight line. Hence, the gompits of the proportions are calculated, interpolated to the conventional mid-points, before being returned to their original scale by means of the application of the anti-gompit. Thus, using an asterisk to denote the shift to conventional midpoints:

$$ {}_5\theta_x^*(i) = \exp\left( -\exp\left( -\begin{bmatrix} 0.4.\left\{-\ln\left(-\ln\left({}_5\theta_{x-5}(i)\right)\right)\right\} \\ +0.6.\left\{-\ln\left(-\ln\left({}_5\theta_x(i)\right)\right)\right\} \end{bmatrix} \right) \right). \quad (1) $$

This can be applied for $x = 20, 25, \ldots, 40$.

To avoid introducing too large a projected component into the projected parity progression ratios, we ignore in all future calculations those estimates of ${}_5\theta_x^*(i)$ which indicate that the proportion of order-specific fertility attained by the mid-point of an age group is less than 0.3. We treat as somewhat speculative calculations based on values of ${}_5\theta_x^*(i) < 0.5$.

***Step 7: Calculate the expected future order increment***

The additional proportion of women reaching each parity in the interval from the mid-point of the age group to the end of childbearing is given by

$${}_5TOFR_{45}(i) - {}_5TOFR_x(i) = {}_5TOFR_{45}(i).\left(1 - {}_5\theta_x^*(i)\right).$$

***Step 8: Derive projected cumulated parity progression ratios and projected parity progression ratios***

The projected cumulated parity progression ratios are calculated by adding the future order increments derived in the previous step to the current cumulated parity progression ratios calculated in Step 1. Thus the proportions of women aged $x$ to $x+5$ *projected* to achieve at least parity $i$ is given by

$${}_5M_x^*(i) = {}_5M_x(i) + {}_5TOFR_{45}(i) \cdot \left(1 - {}_5\theta_x^*(i)\right)$$

and the projected parity progression ratios between parity $i$ and $i+1$ is given by

$${}_5a_x^*(i) = \frac{{}_5M_x^*(i+1)}{{}_5M_x^*(i)}.$$

## INTERPRETATION OF RESULTS

The interpretation of changes in PPRs may be complicated by data errors and should take into account the assumptions underlying the calculations. To the extent that there are errors in the data, care should be taken to not over-interpret the results. PPRs with a substantial projected component must be interpreted with special care, since they may reflect period effects which will cease to apply in the future. If, for example, the AOSFRs reflect a delay in the timing of births around the time that the data are collected and fertility subsequently increases to compensate so that women catch up with the parities achieved by earlier cohorts, then the projected PPRs would be misleading. Equally, a future decline in fertility will reduce the final parities of women even if current rates predict no decline. The projected PPRs for older women are more reliable since they have a smaller projected proportion.

The projected first birth ratio, i.e. ${}_5M_x^*(1)$, is often biased by reference period errors which will not cancel out in the way that they will in the other PPRs. The projected proportions having at least one child will be affected by changes in the timing of entry into motherhood. Since the measure is a synthetic cohort measure derived from current rates, changes in the timing of first births will also affect the cumulated rate and may also make it fall outside the range expected for most real cohorts in developing countries. Short-term fluctuations and tempo effects might mean that the (period) first-birth ratio might occasionally exceed one. This should not be confused with, or interpreted as indicating, a parity progression ratio greater than one, which is definitionally impossible.

The projected first birth ratio may also give an indication of the quality of the data. These rates show the proportion of women expected to become mothers over their reproductive lifetime in a synthetic cohort experiencing these rates. Reference period errors may result in an unlikely projected proportion of women expected to become mothers. Extreme reference period errors might result in this proportion being greater than one. If the cumulated rates are less than 0.9, this is a strong indicator of errors in the data or substantial changes in the timing of first births, although lower proportions may be observed in historical populations with high levels of primary sterility and in contemporary, low-fertility, populations with high levels of voluntary childlessness. A rise in the (projected) proportion of women expected to become mothers, ${}_5PPR_x^{(*)}(0,1)$, has often been observed in populations which have seen a decline in primary sterility.

High-order PPRs may be subject to a large degree of fluctuation as a result of sampling errors in the AOSFRs, since they are derived from relatively small numbers of births. When large sampling errors are present in the data due to small numbers of women, it is acceptable to 'condense' or average two or more adjacent projected PPRs to see trends more clearly.

Examination of the trends in parity progression ratios across cohorts (projected for the 40–44 cohort and younger and actual for the 45–49 cohort and older) can cast further light on fertility trends. A decline in a (projected) PPR indicates a decline in the proportion of women progressing to the following birth and, all other things being equal,

indicates declining fertility. Changes in the PPRs at different parities may result from various factors, with some ratios being affected more than others. Which PPRs change may be a useful guide to fertility changes in the population. For example, the spread of family limitation through parity-specific contraception will be reflected in the decline of PPRs of middle-order parities (say, 3 to 6). A change in higher-order PPRs often reflects parity-independent fertility changes, such as a change in the duration of breast feeding.

## WORKED EXAMPLE

The example analyses data from the 1998 Cambodia Census, which was conducted on 3 March 1998. The method has been implemented in an Excel workbook (see website).

***Step 1: Extract a tabulation of children ever born by age group of mother for each age group with data***

The data from the census, uncorrected, are shown in Table 8.1. The data show a typical pattern of average parities increasing with age up to the 45–49 age group. Thereafter, average parities increase by only a small amount. The data indicate, therefore, that either women older than 50 have omitted some of their children ever born, or that fertility was roughly constant among women finishing their childbearing in the last 10–15 years.

***Step 2: Extract a tabulation showing the number of births in the last year, by mother's age and parity $_5B_x(i)$***

Data from the 1998 Cambodia Census on the number of births reported born in the year before the census, by mother's age group and parity, are shown in Table 8.2.

The data suggest Total Fertility of about 3.25 children per

*TABLE 8.1 Number of women by age group and parity, Cambodia, 1998 Census*

| | Age group | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Parity ($i$) | 15–19 | 20–24 | 25–29 | 30–34 | 35–39 | 40–44 | 45–49 | 50–54 | 55–59 | 60–64 |
| 0 | 635,940 | 200,620 | 105,350 | 55,300 | 33,900 | 22,970 | 16,430 | 9,790 | 7,640 | 6,880 |
| 1 | 30,340 | 94,810 | 79,570 | 38,780 | 21,150 | 16,570 | 13,460 | 9,600 | 8,020 | 7,450 |
| 2 | 5,980 | 62,580 | 113,640 | 66,750 | 33,320 | 21,920 | 17,220 | 12,840 | 9,690 | 8,020 |
| 3 | 1,650 | 21,680 | 88,750 | 75,660 | 45,680 | 27,750 | 20,250 | 15,170 | 11,180 | 8,120 |
| 4 | 1,200 | 6,890 | 44,480 | 72,370 | 56,100 | 35,450 | 24,530 | 18,090 | 13,350 | 10,160 |
| 5 | 630 | 2,060 | 16,580 | 51,600 | 57,050 | 39,100 | 26,660 | 19,660 | 16,100 | 12,040 |
| 6 | 0 | 1,120 | 5,630 | 30,250 | 50,960 | 39,180 | 27,540 | 19,860 | 16,180 | 12,250 |
| 7 | 0 | 510 | 2,110 | 13,390 | 35,780 | 33,190 | 25,850 | 18,880 | 16,170 | 13,690 |
| 8 | 0 | 480 | 1,070 | 5,970 | 20,580 | 24,920 | 23,180 | 16,940 | 14,830 | 11,970 |
| 9 | 0 | 0 | 440 | 1,820 | 9,590 | 15,860 | 16,130 | 12,680 | 11,360 | 9,200 |
| 10 | 0 | 0 | 240 | 880 | 4,840 | 10,420 | 12,790 | 10,720 | 9,800 | 7,750 |
| 11 | 0 | 0 | 120 | 320 | 1,760 | 4,430 | 6,670 | 5,370 | 4,850 | 3,970 |
| 12 | 0 | 0 | 80 | 150 | 760 | 3,220 | 5,350 | 4,810 | 4,780 | 4,230 |
| 13 | 0 | 0 | 40 | 60 | 250 | 990 | 1,960 | 1,790 | 1,790 | 1,440 |
| 14 | 0 | 0 | 0 | 40 | 170 | 620 | 1,020 | 970 | 760 | 540 |
| 15 | 0 | 0 | 0 | 10 | 10 | 250 | 510 | 460 | 440 | 490 |
| 16 | 0 | 0 | 0 | 10 | 30 | 130 | 230 | 160 | 140 | 210 |
| 17 | 0 | 0 | 0 | 0 | 10 | 40 | 80 | 50 | 70 | 70 |
| 18 | 0 | 0 | 0 | 10 | 10 | 30 | 30 | 30 | 40 | 40 |
| TOTAL | 675,740 | 390,750 | 458,100 | 413,370 | 371,950 | 297,040 | 239,890 | 177,870 | 147,190 | 118,520 |
| Average parity | 0.082 | 0.862 | 1.965 | 3.150 | 4.369 | 5.086 | 5.566 | 5.758 | 5.974 | 5.969 |

woman, substantially less than the cohort fertility (5.566) of women aged 45–49. In turn, this suggests either that a dramatic fertility decline is underway in the country, or systematic omission of recent births in the census.

***Step 3: Derive proportions at each attained parity and parity progression ratios***

Proportions attaining parity *i* or higher are calculated from the data in Table 8.1, and are presented in Table 8.3. Parities 12 and greater are combined.

The proportion of women who have had *i* or more births is shown in Table 8.3. Thus, while 80.4 per cent of women aged 45–49 have had at least three births, only 71.9 per cent of women have had four or more births. The associated parity progression ratios, ${}_5a_x(i)$, are shown in Table 8.4.

The ratios suggest that around 93.2 per cent of women aged 45–49 in the 1998 Cambodia Census had had at least one child, while 89.5 per cent of women in this age group who had had three children went on to have a fourth.

***Step 4: Derive age-order specific fertility rates***

The reported births in the last year (by age and parity) in Table 8.2 are divided by the total number of women in each age group (the total, by age group, in Table 8.1). The resulting age-order specific fertility rates are shown in Table 8.5. Thus, for example, ${}_5AOSFR_{25}(1) = 11{,}500/458{,}100 = 0.0251$.

***Step 5: Derive cumulated order-specific fertility rates***

From the data presented in Table 8.5, cumulated age-order specific fertility rates can be derived by summing along each row to the desired age group. Thus, for example, the cumulated age-order specific rate up to the end of the 25–29 age group for parity 3 would be $5\times(0.0006 + 0.0214 + 0.0494) = 0.3572$.

***Step 6: Calculate the age distribution of order-specific fertility rates, and interpolate to conventional ages***

Using the results from Step 5, the proportion of order-specific fertility achieved by the end of each age group is

*TABLE 8.2 Births reported in the year before the census by mother's age group and parity, Cambodia, 1998 Census*

| | Age group | | | | | | |
|---|---|---|---|---|---|---|---|
| Parity (*i*) | 15–19 | 20–24 | 25–29 | 30–34 | 35–39 | 40–44 | 45–49 |
| 0 | | | | | | | |
| 1 | 13,100 | 23,490 | 11,500 | 3,170 | 1,040 | 320 | 190 |
| 2 | 2,340 | 19,180 | 21,320 | 7,830 | 2,010 | 730 | 150 |
| 3 | 410 | 8,370 | 22,640 | 11,470 | 3,790 | 850 | 230 |
| 4 | 210 | 2,430 | 14,330 | 14,110 | 5,200 | 1,450 | 300 |
| 5 | 100 | 600 | 6,220 | 12,440 | 6,740 | 1,830 | 480 |
| 6 | 0 | 280 | 2,020 | 8,360 | 8,120 | 2,870 | 560 |
| 7 | 0 | 90 | 750 | 4,430 | 7,220 | 3,240 | 840 |
| 8 | 0 | 90 | 360 | 2,100 | 4,930 | 3,020 | 750 |
| 9 | 0 | 0 | 90 | 640 | 2,800 | 2,310 | 690 |
| 10 | 0 | 0 | 40 | 300 | 1,390 | 1,570 | 640 |
| 11 | 0 | 0 | 30 | 160 | 630 | 810 | 490 |
| 12 | 0 | 0 | 10 | 40 | 200 | 480 | 320 |
| 13 | 0 | 0 | 10 | 20 | 80 | 250 | 140 |
| 14 | | 0 | 0 | 10 | 60 | 100 | 70 |
| 15 | | | 0 | 0 | 0 | 50 | 30 |
| 16 | | | | 0 | 10 | 10 | 30 |
| 17 | | | | | 0 | 0 | 10 |
| 18 | | | | 0 | 0 | 0 | 10 |
| ASFRs | 0.024 | 0.140 | 0.173 | 0.157 | 0.119 | 0.067 | 0.025 |

derived by dividing the cumulated order-specific fertility rates by the total order specific fertility rate shown in the last column of Table 8.5. The results are shown in Table 8.6.

The proportion of fertility achieved by age 49.5 is assumed to be one. Using the same example as in the previous step, the proportion of third order-specific fertility achieved by age 29.5 is 0.3572/0.5660 = 0.6311.

Because mothers' ages are classified by their age at the census and the average parities derived in Step 1 apply to ages 17½, 22½ etc., to render the two sets of data comparable, we interpolate between pairs of values to get the estimated proportions at ages 22½, 27½ etc. The formula for doing this is given in Equation 1, and the results are shown in Table 8.7.

*TABLE 8.3 Proportion of women ever-attaining each parity by age group, Cambodia, 1998 Census*

| | Age group | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Parity (*i*) | 15–19 | 20–24 | 25–29 | 30–34 | 35–39 | 40–44 | 45–49 | 50–54 | 55–59 | 60–64 |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 0.0589 | 0.4866 | 0.7700 | 0.8662 | 0.9089 | 0.9227 | 0.9315 | 0.9450 | 0.9481 | 0.9420 |
| 2 | 0.0140 | 0.2439 | 0.5963 | 0.7724 | 0.8520 | 0.8669 | 0.8754 | 0.8910 | 0.8936 | 0.8791 |
| 3 | 0.0051 | 0.0838 | 0.3483 | 0.6109 | 0.7624 | 0.7931 | 0.8036 | 0.8188 | 0.8278 | 0.8114 |
| 4 | 0.0027 | 0.0283 | 0.1545 | 0.4279 | 0.6396 | 0.6997 | 0.7192 | 0.7335 | 0.7518 | 0.7429 |
| 5 | 0.0009 | 0.0107 | 0.0574 | 0.2528 | 0.4888 | 0.5803 | 0.6169 | 0.6318 | 0.6611 | 0.6572 |
| 6 | 0.0000 | 0.0054 | 0.0212 | 0.1280 | 0.3354 | 0.4487 | 0.5058 | 0.5213 | 0.5517 | 0.5556 |
| 7 | 0.0000 | 0.0025 | 0.0090 | 0.0548 | 0.1984 | 0.3168 | 0.3910 | 0.4096 | 0.4418 | 0.4522 |
| 8 | 0.0000 | 0.0012 | 0.0043 | 0.0224 | 0.1022 | 0.2051 | 0.2833 | 0.3035 | 0.3320 | 0.3367 |
| 9 | 0.0000 | 0.0000 | 0.0020 | 0.0080 | 0.0469 | 0.1212 | 0.1866 | 0.2082 | 0.2312 | 0.2357 |
| 10 | 0.0000 | 0.0000 | 0.0010 | 0.0036 | 0.0211 | 0.0678 | 0.1194 | 0.1370 | 0.1540 | 0.1581 |
| 11 | 0.0000 | 0.0000 | 0.0005 | 0.0015 | 0.0081 | 0.0327 | 0.0661 | 0.0767 | 0.0874 | 0.0927 |
| 12+ | 0.0000 | 0.0000 | 0.0003 | 0.0007 | 0.0033 | 0.0178 | 0.0383 | 0.0465 | 0.0545 | 0.0592 |

*TABLE 8.4 Observed parity progression ratios by mother's age group and parity, Cambodia, 1998 Census*

| | Age group | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Parity (*i*) | 15–19 | 20–24 | 25–29 | 30–34 | 35–39 | 40–44 | 45–49 | 50–54 | 55–59 | 60–64 |
| 0 | 0.0589 | 0.4866 | 0.7700 | 0.8662 | 0.9089 | 0.9227 | 0.9315 | 0.9450 | 0.9481 | 0.9420 |
| 1 | 0.2377 | 0.5013 | 0.7744 | 0.8917 | 0.9374 | 0.9395 | 0.9398 | 0.9429 | 0.9425 | 0.9333 |
| 2 | 0.3679 | 0.3435 | 0.5840 | 0.7909 | 0.8949 | 0.9149 | 0.9180 | 0.9190 | 0.9263 | 0.9230 |
| 3 | 0.5259 | 0.3378 | 0.4437 | 0.7004 | 0.8389 | 0.8822 | 0.8950 | 0.8958 | 0.9082 | 0.9156 |
| 4 | 0.3443 | 0.3770 | 0.3717 | 0.5909 | 0.7642 | 0.8294 | 0.8578 | 0.8613 | 0.8794 | 0.8846 |
| 5 | 0.0000 | 0.5060 | 0.3698 | 0.5063 | 0.6862 | 0.7732 | 0.8199 | 0.8251 | 0.8345 | 0.8454 |
| 6 | | 0.4692 | 0.4214 | 0.4283 | 0.5915 | 0.7060 | 0.7730 | 0.7858 | 0.8008 | 0.8140 |
| 7 | | 0.4848 | 0.4854 | 0.4091 | 0.5151 | 0.6473 | 0.7244 | 0.7409 | 0.7513 | 0.7446 |
| 8 | | 0.0000 | 0.4623 | 0.3560 | 0.4586 | 0.5909 | 0.6589 | 0.6862 | 0.6965 | 0.7001 |
| 9 | | | 0.5217 | 0.4485 | 0.4498 | 0.5593 | 0.6397 | 0.6577 | 0.6662 | 0.6707 |
| 10 | | | 0.5000 | 0.4054 | 0.3827 | 0.4824 | 0.5534 | 0.5599 | 0.5677 | 0.5864 |
| 11 | | | 0.5000 | 0.4667 | 0.4133 | 0.5438 | 0.5792 | 0.6063 | 0.6232 | 0.6388 |

Due to the inherent uncertainty surrounding the projection of future parity progression, cells in Table 8.7 with a value of less than 0.3 (i.e., less than 30 per cent of order fertility has been achieved by the age indicated) are tinted in grey, and will be excluded from further calculations. Values greater than 0.3 but less than 0.5 are marked in italic, and should be treated with circumspection in further calculations. Calculations depending on these values will be presented in brackets in subsequent tabulations.

***TABLE 8.5*** *Age-order specific and total order fertility rates (TOFR), Cambodia, 1998 Census*

| | Age group | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Parity (*i*)** | **15–19** | **20–24** | **25–29** | **30–34** | **35–39** | **40–44** | **45–49** | **TOFR(i)** |
| 0 | | | | | | | | |
| 1 | 0.0194 | 0.0601 | 0.0251 | 0.0077 | 0.0028 | 0.0011 | 0.0008 | 0.5847 |
| 2 | 0.0035 | 0.0491 | 0.0465 | 0.0189 | 0.0054 | 0.0025 | 0.0006 | 0.6326 |
| 3 | 0.0006 | 0.0214 | 0.0494 | 0.0277 | 0.0102 | 0.0029 | 0.0010 | 0.5660 |
| 4 | 0.0003 | 0.0062 | 0.0313 | 0.0341 | 0.0140 | 0.0049 | 0.0013 | 0.4603 |
| 5 | 0.0001 | 0.0015 | 0.0136 | 0.0301 | 0.0181 | 0.0062 | 0.0020 | 0.3582 |
| 6 | 0.0000 | 0.0007 | 0.0044 | 0.0202 | 0.0218 | 0.0097 | 0.0023 | 0.2959 |
| 7 | 0.0000 | 0.0002 | 0.0016 | 0.0107 | 0.0194 | 0.0109 | 0.0035 | 0.2320 |
| 8 | 0.0000 | 0.0002 | 0.0008 | 0.0051 | 0.0133 | 0.0102 | 0.0031 | 0.1632 |
| 9 | 0.0000 | 0.0000 | 0.0002 | 0.0015 | 0.0075 | 0.0078 | 0.0029 | 0.0996 |
| 10 | 0.0000 | 0.0000 | 0.0001 | 0.0007 | 0.0037 | 0.0053 | 0.0027 | 0.0625 |
| 11 | 0.0000 | 0.0000 | 0.0001 | 0.0004 | 0.0017 | 0.0027 | 0.0020 | 0.0346 |
| 12+ | 0.0000 | 0.0000 | 0.0000 | 0.0002 | 0.0009 | 0.0030 | 0.0025 | 0.0335 |
| ASFR | 0.0239 | 0.1396 | 0.1731 | 0.1574 | 0.1189 | 0.0670 | 0.0247 | 3.5231 |

***TABLE 8.6*** *Proportion of order-specific fertility achieved by the end of each group, Cambodia, 1998 Census*

| | Age group | | | | | | |
|---|---|---|---|---|---|---|---|
| **Parity (*i*)** | **19.5** | **24.5** | **29.5** | **34.5** | **39.5** | **44.5** | **49.5** |
| 0 | | | | | | | |
| 1 | 0.1658 | 0.6799 | 0.8945 | 0.9601 | 0.9840 | 0.9932 | 1.0000 |
| 2 | 0.0274 | 0.4153 | 0.7832 | 0.9329 | 0.9756 | 0.9951 | 1.0000 |
| 3 | 0.0054 | 0.1946 | 0.6311 | 0.8762 | 0.9663 | 0.9915 | 1.0000 |
| 4 | 0.0034 | 0.0709 | 0.4107 | 0.7815 | 0.9334 | 0.9864 | 1.0000 |
| 5 | 0.0021 | 0.0235 | 0.2130 | 0.6331 | 0.8861 | 0.9721 | 1.0000 |
| 6 | 0.0000 | 0.0121 | 0.0866 | 0.4284 | 0.7973 | 0.9606 | 1.0000 |
| 7 | 0.0000 | 0.0050 | 0.0402 | 0.2712 | 0.6895 | 0.9245 | 1.0000 |
| 8 | 0.0000 | 0.0071 | 0.0311 | 0.1868 | 0.5928 | 0.9042 | 1.0000 |
| 9 | 0.0000 | 0.0000 | 0.0099 | 0.0876 | 0.4654 | 0.8556 | 1.0000 |
| 10 | 0.0000 | 0.0000 | 0.0070 | 0.0650 | 0.3639 | 0.7866 | 1.0000 |
| 11 | 0.0000 | 0.0000 | 0.0095 | 0.0654 | 0.3103 | 0.7046 | 1.0000 |
| 12+ | 0.0000 | 0.0000 | 0.0065 | 0.0318 | 0.1724 | 0.6201 | 1.0000 |

***Step 7: Calculate the expected future order increment***

The additional proportion of women reaching each parity in the interval from the mid-point of the age group to the end of childbearing is given by

$${}_5TOFR_{45}(i) - {}_5TOFR_x(i) = {}_5TOFR_{45}(i).\left(1 - {}_5\theta_x^*(i)\right)$$

as shown in Table 8.8.

Thus, the anticipated future order increment after age 30 at parity 3 is given by $0.5660 \times (1 - 0.4655) = 0.3025$. Values shown in brackets are less reliable as they are based on proportions of age-order specific fertility attained by the mid-point of that age group of less than half.

*TABLE 8.7 Proportion of order fertility achieved by the mid-point of each age interval, Cambodia, 1998 Census*

| | Age group | | | | | |
|---|---|---|---|---|---|---|
| Parity (*i*) | 22.5 | 27.5 | 32.5 | 37.5 | 42.5 | 47.5 |
| 0 | | | | | | |
| 1 | *0.4897* | 0.8326 | 0.9409 | 0.9769 | 0.9904 | 0.9999 |
| 2 | 0.2135 | 0.6652 | 0.8915 | 0.9634 | 0.9906 | 0.9999 |
| 3 | 0.0739 | *0.4655* | 0.8044 | 0.9428 | 0.9852 | 0.9999 |
| 4 | 0.0275 | 0.2526 | 0.6624 | 0.8916 | 0.9742 | 0.9998 |
| 5 | 0.0102 | 0.1104 | *0.4751* | 0.8139 | 0.9506 | 0.9998 |
| 6 | | 0.0452 | 0.2738 | 0.6811 | 0.9228 | 0.9997 |
| 7 | | 0.0197 | 0.1539 | 0.5410 | 0.8640 | 0.9996 |
| 8 | | 0.0183 | 0.1061 | *0.4345* | 0.8232 | 0.9996 |
| 9 | | | 0.0430 | 0.2965 | 0.7449 | 0.9995 |
| 10 | | | 0.0311 | 0.2221 | 0.6527 | 0.9994 |
| 11 | | | 0.0341 | 0.1937 | 0.5671 | 0.9993 |
| 12+ | | | 0.0181 | 0.1001 | *0.4473* | 0.9993 |

*TABLE 8.8 Additional proportion attaining parity (i) by the end of childbearing by age group of mother, Cambodia, 1998 Census*

| | Age group | | | | |
|---|---|---|---|---|---|
| Parity (*i*) | 22.5 | 27.5 | 32.5 | 37.5 | 42.5 |
| 0 | | | | | |
| 1 | (0.2984) | 0.0979 | 0.0346 | 0.0135 | 0.0056 |
| 2 | | 0.2118 | 0.0686 | 0.0232 | 0.0059 |
| 3 | | (0.3025) | 0.1107 | 0.0324 | 0.0084 |
| 4 | | | 0.1554 | 0.0499 | 0.0119 |
| 5 | | | (0.1880) | 0.0666 | 0.0177 |
| 6 | | | | 0.0944 | 0.0228 |
| 7 | | | | 0.1065 | 0.0316 |
| 8 | | | | (0.0923) | 0.0289 |
| 9 | | | | | 0.0254 |
| 10 | | | | | 0.0217 |
| 11 | | | | | 0.0150 |
| 12+ | | | | | (0.0185) |

*TABLE 8.9 Projected proportions expected to attain each parity, Cambodia, 1998 Census*

| | Age group | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Parity (*i*) | 20–24 | 25–29 | 30–34 | 35–39 | 40–44 | 45–49 | 50–54 | 55–59 | 60–64 |
| 0 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 1 | (0.7850) | 0.8679 | 0.9008 | 0.9224 | 0.9283 | 0.9315 | 0.9450 | 0.9481 | 0.9420 |
| 2 | | 0.8081 | 0.8410 | 0.8752 | 0.8728 | 0.8754 | 0.8910 | 0.8936 | 0.8791 |
| 3 | | (0.6508) | 0.7216 | 0.7948 | 0.8014 | 0.8036 | 0.8188 | 0.8278 | 0.8114 |
| 4 | | | 0.5833 | 0.6895 | 0.7115 | 0.7192 | 0.7335 | 0.7518 | 0.7429 |
| 5 | | | (0.4408) | 0.5554 | 0.5980 | 0.6169 | 0.6318 | 0.6611 | 0.6572 |
| 6 | | | | 0.4298 | 0.4715 | 0.5058 | 0.5213 | 0.5517 | 0.5556 |
| 7 | | | | 0.3049 | 0.3483 | 0.3910 | 0.4096 | 0.4418 | 0.4522 |
| 8 | | | | (0.1945) | 0.2339 | 0.2833 | 0.3035 | 0.3320 | 0.3367 |
| 9 | | | | | 0.1466 | 0.1866 | 0.2082 | 0.2312 | 0.2357 |
| 10 | | | | | 0.0895 | 0.1194 | 0.1370 | 0.1540 | 0.1581 |
| 11 | | | | | 0.0477 | 0.0661 | 0.0767 | 0.0874 | 0.0927 |
| 12+ | | | | | (0.0363) | 0.0383 | 0.0465 | 0.0545 | 0.0592 |

***Step 8: Derive expected projected proportions attaining each parity order and projected parity progression ratios***

The projected cumulated parity progression ratios are calculated by adding the future order increments presented in Table 8.8 to the proportion of women who have ever attained each parity calculated in Step 3 and presented in Table 8.3.

The projected PPRs for cohorts who are still bearing children can now be compared directly with the completed PPRs of women aged 50 or more. The results are shown in Table 8.9.

In the 25–29 age group, the projected proportion with parity 3 and over is 0.3483 (from Table 8.3)+ 0.3025 (from Table 8.8) = 0.6508. Values shown in brackets are less reliable as they are based on proportions of age-order specific fertility attained by the mid-point of that age group of less than half.

Projected parity progression ratios are derived by division of the proportions expected to attain each successive parity in any given age cohort (Table 8.10).

**Projected parity progression ratios**

The parity progression ratio from 0 to 1 shows the proportion of women who have ever had a child. Note how the projected parity progression ratio at this parity suffers from extensive bias, as discussed in detail below. The proportion of childless women has remained roughly constant by cohort. However, as can be seen from Figure 8.1, at parities 2 and greater there is a clear trend for slightly lower parity progression in each successively younger cohort.

The projected proportions expected to attain each parity from the 1998 Cambodia Census are shown in Figure 8.2.

The effects of the lower parity progression ratios for younger cohorts is evident in the proportions expected to attain each parity. Thus while 55 per cent of women aged 60–64 had had 6 or more children, in the cohort now completing its childbearing (that aged 45–49), the proportion had declined to just over half (50.6 per cent).

The data indicate the makings of an incipient decline in fertility that began roughly – when the cohort currently aged 50–54 was having their children, approximately 25–30 years before the census.

**Entry into motherhood**

In the case of the Cambodian data, the actual proportion of women 45–49 who had ever had a child was 0.9315. The projected proportions, by current age of mother, are shown in the first row of Table 8.10.

It would certainly appear that there have been some significant changes in the timing of entry into motherhood among younger cohorts, although the results for women in their twenties, and quite probably those aged 30–34 too, are implausible for reasons discussed in relation to the derivation of the results earlier. The projected parity

*TABLE 8.10 Projected parity progression ratios, Cambodia, 1998 Census*

| | Age group | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Parity (*i*) | 20–24 | 25–29 | 30–34 | 35–39 | 40–44 | 45–49 | 50–54 | 55–59 | 60–64 |
| 0 | (0.7850) | 0.8679 | 0.9008 | 0.9224 | 0.9283 | 0.9315 | 0.9450 | 0.9481 | 0.9420 |
| 1 | | 0.9312 | 0.9337 | 0.9488 | 0.9403 | 0.9398 | 0.9429 | 0.9425 | 0.9333 |
| 2 | | (0.8053) | 0.8580 | 0.9081 | 0.9182 | 0.9180 | 0.9190 | 0.9263 | 0.9230 |
| 3 | | | 0.8083 | 0.8676 | 0.8878 | 0.8950 | 0.8958 | 0.9082 | 0.9156 |
| 4 | | | (0.7558) | 0.8055 | 0.8404 | 0.8578 | 0.8613 | 0.8794 | 0.8846 |
| 5 | | | | 0.7738 | 0.7885 | 0.8199 | 0.8251 | 0.8345 | 0.8454 |
| 6 | | | | 0.7094 | 0.7388 | 0.7730 | 0.7858 | 0.8008 | 0.8140 |
| 7 | | | | (0.6379) | 0.6715 | 0.7244 | 0.7409 | 0.7513 | 0.7446 |
| 8 | | | | | 0.6266 | 0.6589 | 0.6862 | 0.6965 | 0.7001 |
| 9 | | | | | 0.6104 | 0.6397 | 0.6577 | 0.6662 | 0.6707 |
| 10 | | | | | 0.5326 | 0.5534 | 0.5599 | 0.5677 | 0.5864 |
| 11 | | | | | (0.7611) | 0.5792 | 0.6063 | 0.6232 | 0.6388 |

*FIGURE 8.1* *Projected parity progression ratios by parity and age group, Cambodia, 1998 Census*

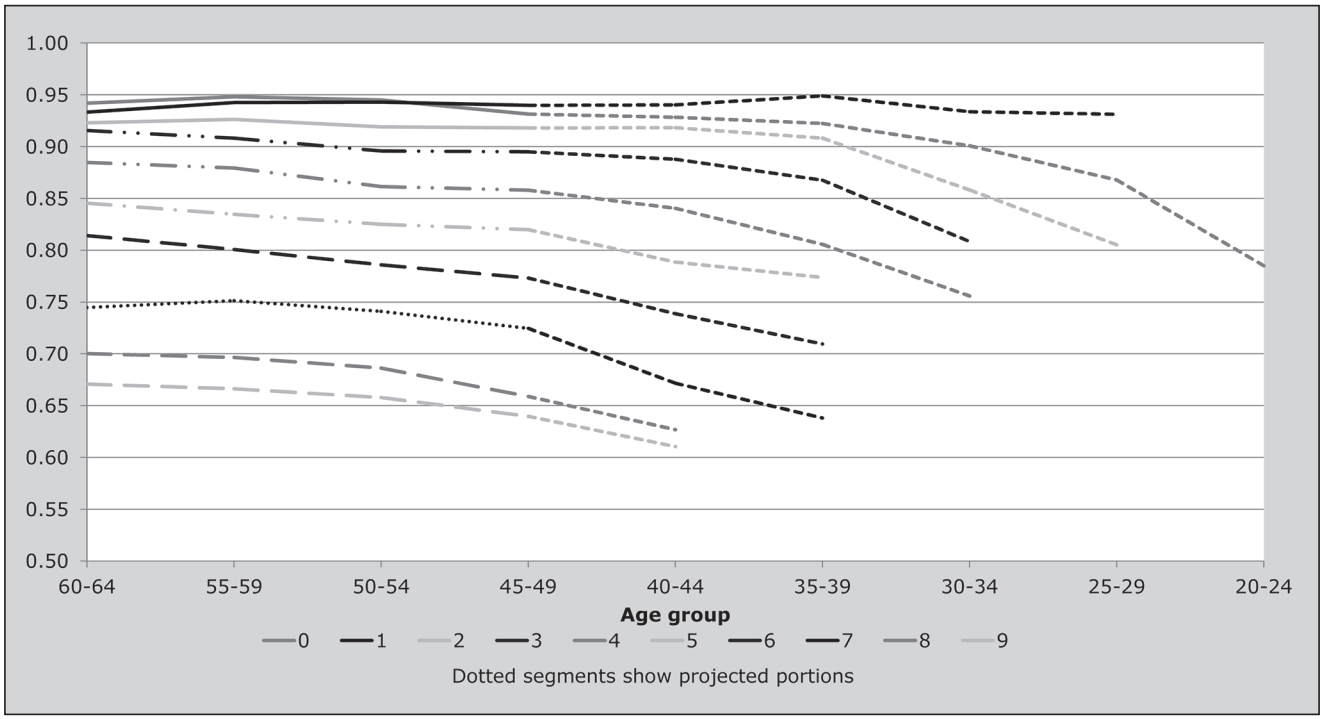
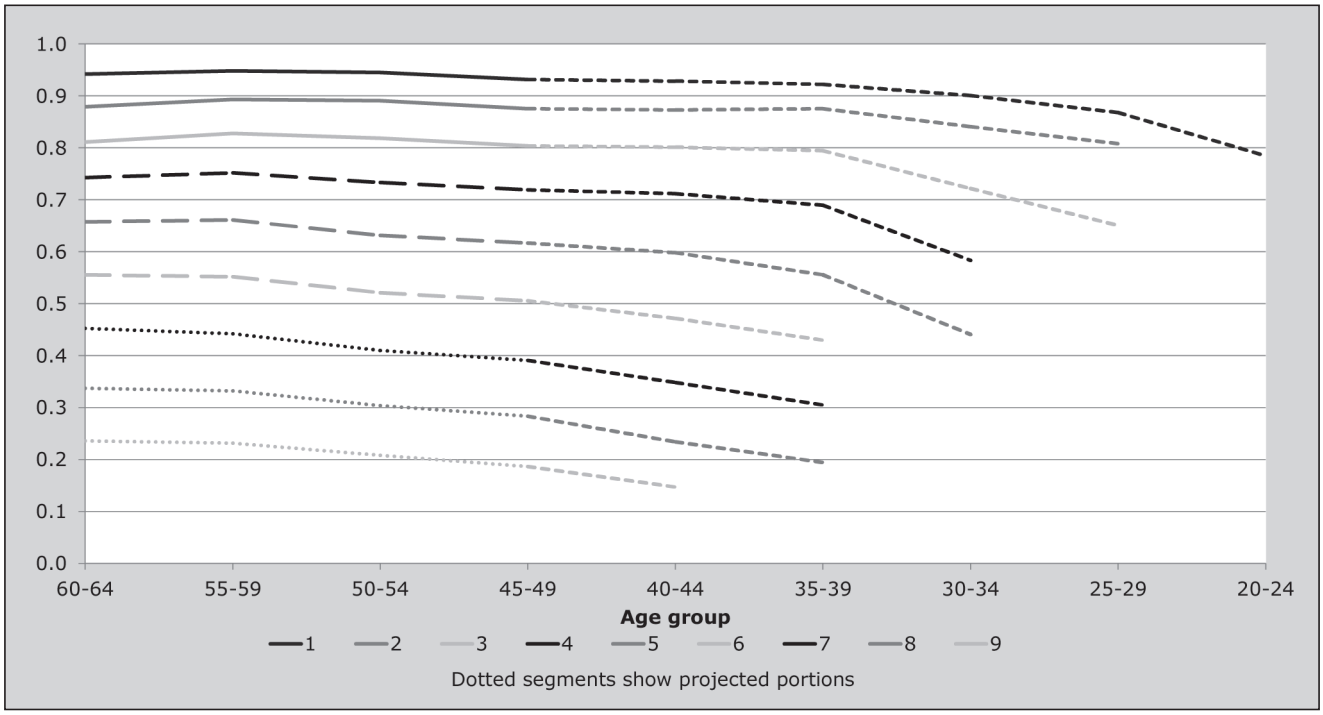



*FIGURE 8.2* *Projected proportion attaining each parity by age group, Cambodia, 1998 Census*



progression ratios for women in their twenties are strongly affected both by changes in the timing of childbearing, as well as being strongly affected by the projected components. These results should be ignored, or alternatively a life-table approach to estimating entry into motherhood should be applied. Such methods and their calculation are outside the scope of this manual.

With the exception of this metric, however, the calculation of projected PPRs should eliminate the effect of reference period errors in the resulting PPRs in so far as they are independent of parity. This can be seen in the estimates of the projected parity progression ratios for other parities.

The data for the 20–24 to 40–44 cohorts contain a projected component which is more significant the younger the cohort. Data for older cohorts are based on completed fertility. The projected parity progression ratios for women aged 30–44 indicate a systematic decline in the proportion of women in each successively younger cohort expected to progress to the next parity by the end of their childbearing years. This is indicative of fertility decline underway.

## FURTHER READING AND REFERENCES

A similar approach to that presented above, but which makes use of the detailed data collected in birth histories and also takes more careful account of the censoring and selection effects is described in Brass and Juárez (1983).

There is a sizable literature on the use and problems associated with period (as opposed to cohort) parity progression ratios. These fall outside of the scope of this manual. The interested reader is referred to Whelpton (1946), Feeney and Yu (1987), Ní Bhrolcháin (1987), Hinde (1998), and Bongaarts and Feeney (1998).


Bongaarts J and G Feeney. 1998. "On the quantum and tempo of fertility", *Population and Development Review* **24**(2):271–291. doi: 10.2307/2807974

Brass W. 1985. *Advances in Methods for Estimating Fertility and Mortality from Limited and Defective Data.* London: Centre for Population Studies, London School of Hygiene and Tropical Medicine.

Brass W and F Juárez. 1983. "Censored cohort parity progression ratios from birth histories", *Asian and Pacific Census Forum* **10**(1):5–12. http://hdl.handle.net/10125/3583

Feeney G and J Yu. 1987. "Period parity progression measures of fertility in China", *Population Studies* **41**(1):77–102. doi: 10.1080/0032472031000142546

Hinde A. 1998. *Demographic Methods.* London: Arnold.

Ní Bhrolcháin M. 1987. "Period parity progression ratios and birth intervals in England and Wales, 1941–1971: A synthetic life table analysis", *Population Studies* **41**(1):103–125. doi: 10.1080/0032472031000142556

Preston SH, P Heuveline and M Guillot. 2001. *Demography: Measuring and Modelling Population Processes.* Oxford: Blackwell.

Whelpton PK. 1946. "Reproduction rates adjusted for age, parity, fecundity, and marriage", *Journal of the American Statistical Association* **41**(236):501–516. doi: 10.1080/01621459.1946.10501893

# Chapter 9 Estimation of fertility by reverse survival

*Ian M Timæus and Tom A Moultrie*

## DESCRIPTION OF METHOD

Reverse survival is a method for estimating fertility from data collected in a census or single-round survey that can be used even if no questions have been asked about fertility directly. In a population closed to migration, the population of any age $x$ are the survivors of the births in that population $x$ completed years previously. This implies that the number of births occurring $x$ years ago can be calculated, provided that one can estimate the life table survival probabilities from birth to age $x$ (that is, $\frac{L_x}{l_0}$).

'Reverse surviving' the population to its birth year and dividing by an estimate of the total population in that year gives the crude birth rate, while dividing by an estimate of women of childbearing age gives the General Fertility Ratio. By combining reverse survival estimates of past births and women according to age with estimates of, or a reasonable assumption about, the age pattern of fertility, one can also estimate Total Fertility.

Asking about births in the last year or the date of women's last live birth provides an estimate only of current fertility. In contrast, reverse survival methods can provide estimates of fertility for the last 15 years. Moreover, unlike fertility estimates from birth histories, which are usually collected only from women aged 15 to 49, fertility estimates produced by reverse survival do not become increasingly truncated at older ages as they are calculated for more distant periods. So long as a single-year age distribution of children is available, the approach can produce an annual series of fertility estimates. In practice though, the data on age collected in developing countries are seldom sufficiently accurate to yield an undistorted time series.

## CAVEATS AND WARNINGS

The methods described here depend heavily on the accuracy of the reported age distribution of the population being investigated. Errors in the age distribution, such as omission of infants and young children and heaping of reports on particular ages, can have a significant impact on the results derived. As *Manual X* notes, "because these types of deficiencies are all too frequently characteristic of the data sets available, reverse survival methods are often ineffective in producing reliable fertility estimates. Their usefulness depends mainly upon the fact that they often provide independent fertility estimates which can be used to assess the plausibility of those obtained by other means" (UN Population Division 1983: 178).

Under-enumeration of infants and young children is common in censuses and surveys in less developed countries. Such under-enumeration of children, especially those aged 2 or less, will result in estimates of fertility that are proportionally too low. Thus, reverse survival methods tend to underestimate current fertility and to produce an exaggerated impression of recent fertility decline.

In addition, inaccurate age reporting, in particular heaping of reports on preferred ages, may result in very erratic time series of estimates of fertility.

Before applying any reverse survival method of estimating fertility, investigations into the quality of the age and sex data should be undertaken. The methods should be applied with caution where the data are deemed to be of poor quality.

Although it is possible, it is generally inadvisable to produce reverse survival estimates of fertility going back more than 15 years before the inquiry concerned. The first reason for this is that both migration and differential under-enumeration are likely to distort estimates based on the age distribution of young adults aged 15 or more. Second, more distant estimates of fertility become increasingly dependent on accurate assessment of the mortality of older women.

Since the results depend on what is assumed about survivorship, they are fairly insensitive to assumptions made about the mortality of either children or women aged less

than 65. For example, if the probability of having died of a child born about 12 years ago was 20 per cent and was under-estimated by a quarter at 16 per cent, then the resulting overestimate of the number of births would be (100–16)/(100–20), which is only 5 per cent. Similarly, even if the child had an older mother, her probability of having died would seldom be much greater than 20 per cent. If this probability was underestimated by a quarter, fertility would be overestimated by only 4.8 per cent. Moreover, if the bias in the mortality estimates is in the same direction for both children and mothers, the resulting errors in the estimates of fertility will partly cancel out.

## METHOD 1: ESTIMATION OF THE CRUDE BIRTH RATE FOR FIVE-YEAR PERIODS

### Data requirements and assumptions

#### *Tabulations of data required*

To derive average birth rates for five-year periods preceding an inquiry, the following data are required:

- enumerated population aged less than age 15, classified by five-year age group;
- total population enumerated either at two points in time or at a single point in time, together with an estimate of the growth rate; and
- estimated life table measures of survivorship for children, ${}_5L_0$, ${}_5L_5$ and ${}_5L_{10}$.

#### *Important assumptions*

The population is assumed to have been closed to migration for as many years as are covered by the reverse survival estimates. However, because children usually migrate with their mothers, errors in the numerator and denominator of the estimated rates largely cancel out. Significant bias will result only if migration flows are large and migrants have different fertility from the rest of the population.

### Application of method

#### *Step 1: Calculate the life table survivorship of children*

In order to reverse survive the population in the three age groups 0–4, 5–9 and 10–14, one needs only values of ${}_5L_0$, ${}_5L_5$, and ${}_5L_{10}$, the person-years lived by the stationary population between birth and age 5, and between ages 5 and 10, and 10 and 15, respectively. The source for these measures might be an empirically-based life table for the population under study or a model life table that is believed to be appropriate (for example, one of the Princeton regional model life tables). Given the high degree of aggregation across ages, one can assume that mortality is constant within each five-year period of time preceding the inquiry. Overestimates of mortality will produce overestimates of fertility (as they imply that the enumerated population represents the survivors of a relatively large cohort of births) and vice versa.

Since mortality changes rapidly in the first year of life, if an empirical life table is used, ${}_5L_0$ should be calculated as ${}_5L_0 = {}_1L_0 + {}_4L_1$. Should the empirical life tables be sex-specific, a combined-sex life table should be derived for ages 0 to 14 by assuming an appropriate sex ratio at birth. This is achieved by means of the formula

$$l_x = \frac{\frac{S_0}{100}.l_x^m + l_x^f}{\frac{100 + S_0}{100}},$$

where $l_x^m$ and $l_x^f$ are the life table values of $l_x$ for boys and girls respectively, and $S_0$ is the sex ratio at birth (expressed as the number of male births per 100 female births).

If a model life table is used to represent mortality in childhood, ${}_5L_0$, ${}_5L_5$, and ${}_5L_{10}$ are obtained by interpolation of appropriate values from the tabulated series of models.

#### *Step 2: Estimate the mid-period populations*

In order to estimate an annual birth rate for up to three preceding periods of five years each, an estimate of the total population at the mid-point of each of the three periods is required. In the absence of detailed information that can be used directly (for example, the results of a census conducted 10 years earlier), and given the inherent loss of precision introduced by aggregating both population counts and mortality into five-year groups, obtaining a precise estimate is unnecessary. The easiest way to estimate the total population $d$ years before an inquiry, $N(t-d)$, where $d = 2.5$, 7.5 and 12.5, is to assume a constant rate of growth, $r$, and apply this to the population enumerated at time $t$, $N(t)$, i.e.

$$N(t-d) = N(t).\exp(-d.r).$$

The growth rate, $r$, can be estimated from information on the size of the total population, $P$, at two points in time, $t_0$ and $t_1$, such that

$$r = \frac{\ln\left(\frac{P_1}{P_0}\right)}{(t_1 - t_0)}.$$

Thus, for example, the population 2½ years before the more recent inquiry is

$$N(t-2.5) = N(t).\exp(-2.5r)$$

and similarly for the estimates of the population at other points in time.

***Step 3: Estimate births in each five-year period preceding the inquiry***

Define $B(t-5,t)$ to be the number of births that occurred each year in the period $t-5$ to $t$ years before an inquiry where $t = 0$, 5 or 10. Define also ${}_5N_x(t)$ to be the enumerated population aged between $x$ and $x+5$ at the date of the inquiry. It follows that the average number of births in each of the five years from $t-5$ to $t$ is

$$B(t-5,t) = {}_5N_0(t).l_0/{}_5L_0$$
$$B(t-10,t-5) = {}_5N_5(t).l_0/{}_5L_5$$

and

$$B(t-15,t-10) = {}_5N_{10}(t).l_0/{}_5L_{10}.$$

The crude birth rate for each period is then calculated by dividing the number of births in each period by the corresponding mid-period populations from Step 2

$$CBR(t-5,t) = \frac{B(t-5,t)}{N(t-2.5)}$$

and similarly for the earlier periods.

**WORKED EXAMPLE**

The basic principles of reverse-survival estimation of fertility are illustrated here using data from the Census conducted in Cambodia in 2008.

***Step 1: Calculate the life table survivorship of children***

Survival factors are derived from the World Health Organization life table for Cambodia for 2000. Values of ${}_nL_x$ for both sexes in this life table are shown in Table 9.1 for a radix of 100,000.

*TABLE 9.1 Extracted values from WHO life table for Cambodia, 2000*

| Age | $n$ | $X$ | $l_x$ | $d_x$ | ${}_nL_x$ |
|---|---|---|---|---|---|
| <1 | 1 | 0 | 100,000 | 7,956 | 94,431 |
| 1–4 | 4 | 1 | 92,044 | 2,684 | 361,734 |
| 5–9 | 5 | 5 | 89,360 | 887 | 444,581 |

Based on these data, one can calculate the following quantities

$$\frac{l_0}{{}_5L_0} = \frac{100{,}000}{94{,}431+361{,}734} = 0.21922$$
$$\frac{l_0}{{}_5L_5} = \frac{100{,}000}{444{,}581} = 0.22493.$$

***Step 2: Estimate the mid-period populations***

Using the samples from Cambodian censuses provided by IPUMS International, the population of Cambodia on 3 March 1998 was 11,412,540 people. By the same date in 2008, it was 13,401,210. The population under the age of 5 in 2008 was 1,736,790 and that aged between 5 and 10 was 1,472,470.

Since the censuses were conducted exactly ten years apart, the intercensal growth rate is calculated as

$$r = \frac{\ln\left(\frac{13401210}{11412540}\right)}{10} = 1.60632 \text{ per cent per annum.}$$

The population 2.5 and 7.5 years before the second census is therefore estimated to be

$$N(t-2.5) = N(t).\exp(-2.5r) = 12{,}873{,}706$$
$$N(t-7.5) = N(t).\exp(-7.5r) = 11{,}880{,}172.$$

***Step 3: Estimate births in each five-year period preceding the inquiry***

The estimated number of births in each of the five years before the census is given by

$$B(t-5,t) = 1{,}736{,}790 \times 0.21922 = 380{,}739.$$

For the five years before that, it is

$$B(t-10,t-5) = 1{,}472{,}471 \times 0.22493 = 331{,}202.$$

Dividing the births by the estimated mid-period population yields crude birth rates for Cambodia of 29.6 births per 1000 in the period 2003–2008, and 27.9 births per 1000 in the period 1998–2003.

These figures are of the right order of magnitude: the US Census Bureau estimates the crude birth rate in the country as 28 in 2000 and 27 in 2005, while the UN Population Division's estimate for 2000–2005 is 25.1 births per 1000. Note, however, that the reverse survival estimates suggest that a slight rise in fertility occurred during the 1990s. This

finding could well be spurious and might result from errors in the initial estimates of the population or an inappropriate choice of life table.

## METHOD 2: ESTIMATION OF GENERAL FERTILITY AND TOTAL FERTILITY RATIOS

The second application of the concept of reverse survival described here produces annual estimates of the General Fertility and Total Fertility Ratios for up to 15 years before the inquiry. To calculate the General Fertility Ratio requires only that one estimates the past size of the population of adult women from the number of women enumerated in an inquiry by allowing for adult mortality. To calculate Total Fertility, however, requires information not only on the number of births occurring each year, but also on the ages of the mothers of these newborn children. One relatively simple way of estimating this information, if it is not known, is to apportion births to age groups of mother using independent estimates of the age distribution of fertility.

### Data requirements and assumptions

#### *Tabulations of data required*

To derive General Fertility Ratios for individual years, the following data are required:

- Tabulations of the population (of both sexes) aged 0 to 14, by single years of age.
- Tabulations of the female population aged 15 to 64 by five-year age group.
- Cohort survival probabilities, $L_x$, for children aged 0 to 14 of both sexes.
- Survivorship ratios, ${}_5L_{x-5}/{}_5L_x$ for adult women for each of the three five-year periods preceding the inquiry.

In respect of the mortality estimates, the implementation of the method in the associated Excel workbook (see website) allows these to be specified either by reference to period-specific parameters $\alpha$ and $\beta$ of appropriate relational model life tables, or to identified values of ${}_5q_0$ (for children) and ${}_{45}q_{15}$ (for adult women) for each of the three five-year periods preceding the enquiry.

To produce estimates of Total Fertility, one also requires either

- A single age-specific fertility distribution that is assumed to apply to the entire period covered by the estimates, or
- Two age-specific fertility distributions, one of which applies to a date reasonably close to the index inquiry and the other to a date approximately 15 years prior to that.

Either a series of fertility rates or the parameters of a relational Gompertz model fitted to a standard fertility schedule can be used as an input to the calculations.

Note that only the estimated shape of the fertility distribution is based on these fertility schedules. It is the estimated number of births relative to the population of women of childbearing age that almost entirely determines the estimates of the General and Total Fertility Ratios.

#### *Important assumptions*

The population is assumed to have been closed to migration for as many years as are covered by the reverse survival estimates. However, because children usually migrate with their mothers, errors in the numerator and denominator of the estimated rates largely cancel out. Significant bias will result only if migration flows are large and migrants have different fertility from the rest of the population.

### Application of method

#### *Step 1: Estimate the number of births in each year before the inquiry*

The enumerated population aged $x$ in any inquiry represents the survivors from the births that occurred in the 12-month period centred on the date $x+0.5$ years before the inquiry. Algebraically,

$$B_{x+0.5} = \frac{N_x}{{}^cL_x}, 0 \le x \le 14.$$

The measure of survivorship used in this calculation, ${}^cLx$, is a *cohort* survival factor. It depends on mortality at successive ages in successive years leading up to the inquiry. Appropriate cohort estimates of mortality may be available from the inquiry used to estimate fertility. Such estimates include indirect estimates from children ever born and surviving, in the case of a census, and direct estimates from a cohort analysis of the birth histories, in the case of a fertility survey. Brass (1979) describes a simple procedure for estimating $L_x$ directly from proportions of children surviving according to their mothers' age group.

If a series of cohort survival ratios is not readily to hand, one can be derived from estimates of period mortality by single years of age for each five-year period before the inquiry.

Working with a relational logit system of model life tables, define $\alpha_T$ and $\beta_T$ to be the parameters that generate a life table for period $T$, where $T=0$ refers to the period 0–4

years before the inquiry, $T = 5$ to the period 5–9 years before the inquiry and $T = 10$ to the period 10–14 years before the inquiry. For a given standard, indexed by the superscript $s$,

$$Y_{x,T} = \alpha_T + \beta_T . Y_x^s$$

where $Y$ is the logit function:

$$Y_x = \frac{1}{2}\ln\left(\frac{1-l(x)}{l(x)}\right) \quad (1)$$

and $l(x)$ (or ${}_xp_0$) refers to the proportion surviving from birth to exact age $x$ in a life table with a radix of 1. Equation 1 implies that

$$l_x = \left(1+\exp\left(2Y_x\right)\right)^{-1}$$

and that

$$l_{x,T} = \left(1+\exp\left(2\left(\alpha_T + \beta_T . Y_x^s\right)\right)\right)^{-1}. \quad (2)$$

At ages other than infancy, one can approximate $L_{x,T}$, the person-years lived between $x$ and $x+1$ in period $T$, by assuming that survivorship declines linearly on the logistic scale and, therefore, that the logit of $L_{x,T}$ is the average of $Y_{x,T}$ and $Y_{x+1,T}$. From Equation 2

$$\begin{aligned} L_{x,T} \approx l_{x+0.5,T} &= \left(1+\exp\left(2\left(\frac{Y_{x,T}+Y_{x+1,T}}{2}\right)\right)\right)^{-1} \\ &= \left(1+\exp\left(Y_{x,T}+Y_{x+1,T}\right)\right)^{-1}, 0 < x \le 14. \end{aligned} \quad (3)$$

The values of $l(x)$ in systems of model life tables are often tabulated by single ages up to age 5, and then at every fifth year of age. If this is the case for children aged 5 or more, one can assume that logit survivorship declines linearly over the entire age range $x$ to $x+5$. Thus, for example, if a life table has tabulated values at $x = 5$ and $x = 10$, the estimated value of $L_{9,T}$ would be given by

$$\begin{aligned} L_{9,T} \approx l_{9.5,T} &= \left(1+\exp\left(2\left(\frac{1}{10}Y_{5,T}+\frac{9}{10}Y_{10,T}\right)\right)\right)^{-1} \\ &= \left(1+\exp\left(0.2Y_{5,T}+1.8Y_{10,T}\right)\right)^{-1}. \end{aligned}$$

For infants, one should allow for the concentration of deaths in the first days and weeks of life. In medium and high mortality populations, one can approximate person years lived in the first year of life as

$$L_{0,T} = 0.3 + 0.7\left(\exp\left(2Y_{1,T}\right)\right)^{-1}. \quad (4)$$

Survivorship ratios from one age to the next, $P_{x,T}$, in time period $T$, are derived from the ratio of successive values of $L_{x,T}$:

$$\begin{aligned} P_{x,T} &= \frac{L_{x,T}}{L_{x-1,T}}, 0 < x \le 14 \\ P_{0,T} &= L_{0,T}. \end{aligned} \quad (5)$$

Once estimates of the survivorship of children by single years of age and five-year time periods have been obtained in either the way just outlined or by some other procedure, an estimate of cohort survivorship by single years of age for single-year age cohorts can be calculated as follows. Recall that $P_{a,T}$ is the survivorship ratio between ages $a$ and $a+1$ in time period $T$ (where $T = 0$, 5 or 10, corresponding to five-year periods 0–4, 5–9 and 10–14 years before the inquiry). Further define $S_{a,t}$ to be the survivorship ratio between ages $a$ and $a+1$ in the period $t$ to $t+1$ years before the inquiry, $0 \le t \le 14$. Using linear interpolation to estimate survivorship for the intermediate years

$$\begin{aligned} S_{a,t} &= P_{a,0} && 0 \le t \le 2 \\ S_{a,t} &= P_{a,T}\left(1-\frac{t-T-2}{5}\right) \\ &\quad + P_{a,T+5}\left(\frac{t-T-2}{5}\right) && 2 < t-T < 8 \\ S_{a,t} &= P_{a,10} && 12 < t < 15. \end{aligned}$$

The proportion of births occurring $x$ to $x+1$ years earlier that survive to the time of the inquiry, ${}^cL_x$, can then be calculated as

$${}^cL_x = S_{0,x}.S_{1,x-1}.\ldots.S_{x-1,1}.S_{x,0}\ . \quad (6)$$

The number of births in each year before an inquiry, centred on the point mid-way through that year (i.e. 6 months before the date of the inquiry), is thus

$$B_{x+0.5} = \frac{N_x}{{}^cL_x}, 0 \le x \le 14$$

where $N_x$ is the number of children aged $x$ reported in the inquiry.

***Step 2: Estimate the mid-year populations of women by five-year age group***

The calculation of survivorship for women aged 15 to 64 at the date of the inquiry can be done in a straightforward way because mortality is usually fairly low in the central adult ages. Thus, even approximate estimates of mortality will enable one to produce a satisfactory estimate of the past population from the enumerated population.

The absolute variation of mortality with age within any five-year age group is small. One can therefore approximate ${}_5L_x$ by linear interpolation between $Y_x$ and $Y_{x+5}$. This means that one can estimate survivorship between five-year age groups at time $T$ (where $T = 5$, 10 and 15) as

$$ {}_5P_{x,t} = \frac{{}_5L_{x+5}}{{}_5L_x} = \frac{1+\exp\left(2\alpha_T + \beta_T\left(Y^s_x + Y^s_{x+5}\right)\right)}{1+\exp\left(2\alpha_T + \beta_T\left(Y^s_{x+5} + Y^s_{x+10}\right)\right)} \ . \quad (7) $$

Starting with the population enumerated at $T = 0$, the number of women in each five-year age group $T + 5$ years before the inquiry can then be calculated from the number at $T$:

$$ {}_5N^f_{x,T+5} = \frac{{}_5N^f_{x+5,T}}{{}_5P_{x,T}},\ 10 \le x < 60,\ T\ =\ 0,\ 5,\ 10\ . \quad (8) $$

Furthermore, since the age structure of a population changes only slowly, the mid-year populations of women in the age groups 10–14, 15–19,…, 60–64 for each year before the date of the inquiry can be estimated by interpolating linearly between the population estimates for 0, 5, 10 and 15 years before the inquiry produced by Equation 8. For example, to estimate the number of women aged 20–24 at the point 8½ years before the inquiry, the formula would be

$$ {}_5N_{20,8.5} = 0.3\left({}_5N_{20,5}\right) + 0.7\left({}_5N_{20,10}\right). $$

***Step 3: Derive General Fertility Ratios***

The General Fertility Ratio for the year centred on the point $x - 0.5$ years before the inquiry is

$$ GFR_{x+0.5} = \frac{B_{x+0.5}}{\sum_{a=15,5}^{45} {}_5N_{a,x+0.5}}, 0 \le x \le 14 \quad (9) $$

where the denominator is the total number of women aged between 15 and 49 at the mid-point of the year during which the births occurred.

***Step 4: Estimate age-specific and Total Fertility***

A natural extension to the calculation of General Fertility Ratios is to make use of a schedule describing the age-pattern of fertility in the population being studied to estimate Total Fertility by a procedure akin to indirect standardization.

Such a fertility schedule might come from data on recent births collected in the same inquiry as is being analysed by reverse survival methods. As only information on the age pattern of fertility is obtained from these data, no need exists to adjust the reports for reference-period errors. If a second fertility schedule is available from a previous census or survey of the same population (preferably conducted about 15 years earlier) one can interpolate between the two schedules, or if necessary extrapolate, to estimate the shape of the fertility schedule in each year for which one intends to estimate Total Fertility.

However, even if one suspects that fertility has been changing, having two schedules is not essential, since the age pattern of fertility evolves only gradually and estimates of Total Fertility are not very sensitive to the exact assumptions made about the shape of the fertility distribution. Thus, a single schedule (preferably referring to about the middle of the estimation period) will suffice.

If fertility has been estimated using a relational Gompertz model, in combination with the chosen standard, the parameters $\alpha$ and $\beta$ derived during the estimation of fertility define the shape of the fertility schedule. Accordingly, values of $\alpha$ and $\beta$ can be used to determine the shape of the fertility schedule(s) used in the estimation of Total Fertility.

Once one has estimated the proportion of Total Fertility occurring in each age group for each year before the inquiry, these proportions can be applied to the population of women in each age group in each year to estimate the number of births that would have occurred to women in that age group if Total Fertility equalled one child per woman.

Thus, once one has selected a fertility schedule (scaled to a Total Fertility of 1), ${}_5f^*_{a,x+0.5}$, for each age group ($a = 15$, 20,…, 45) for each of the 15 years ($x$) before an inquiry, the expected number of births to women in each age group in each year is given by

$$ {}_5B^*_{a,x+0.5} \ =_5 N_{a,x+0.5} \cdot {}_5f^*_{a,x+0.5}, 0 \le x \le 14\ . $$

It follows that the total number of births that would have occurred in year $x$ if Total Fertility had equalled 1 is

$$B^{*}_{x+0.5} = \sum_{a=15}^{45} {}_5N_{a,x+0.5} \cdot {}_5f^{*}_{a,x+0.5}, 0 \le x \le 14 .$$

However, Step 2 yielded an estimate of the actual number of births in each year, $B_{x+0.5}$. The estimate of Total Fertility for each year is thus the ratio of $B$ to $B^{*}$

$$TF_{x+0.5} = \frac{B_{x+0.5}}{B^{*}_{x+0.5}}, 0 \le x \le 14 .$$

Estimates of the age-specific fertility rates for the year can be produced by multiplying the proportional rates, ${}_5f^{*}_{a,x}$, for each age group ($a$) and year ($x$) by the estimate of Total Fertility for that year. It should be noted, however, that the results will merely reproduce the age pattern of fertility that was input into the calculations: this method provides no new information on the age pattern of fertility.

## WORKED EXAMPLE

The worked example, implemented in an Excel workbook (see website), estimates fertility trends in Cambodia from the 2008 Census. Sample microdata are available from IPUMS International and the reference date for the census was 3 March 2008. In decimal notation, this date corresponds to 2008.17.

The numbers of children aged less than 15 by single years of age, and the numbers of women aged 10 to 64 in five-year age groups extracted from the data are shown in Table 9.2.

*TABLE 9.2 Numbers of children by single year of age, and women by five-year age group, Cambodia, 2008 Census*

| Age | Children | Age group | Women |
|---|---|---|---|
| 0 | 281,260 | 10–14 | 815,930 |
| 1 | 261,320 | 15–19 | 780,320 |
| 2 | 268,410 | 20–24 | 697,160 |
| 3 | 286,810 | 25–29 | 626,430 |
| 4 | 278,990 | 30–34 | 361,650 |
| 5 | 293,760 | 35–39 | 435,880 |
| 6 | 293,490 | 40–44 | 393,760 |
| 7 | 302,060 | 45–49 | 352,520 |
| 8 | 315,970 | 50–54 | 294,280 |
| 9 | 267,190 | 55–59 | 230,200 |
| 10 | 326,980 | 60–64 | 160,590 |
| 11 | 280,260 | | |
| 12 | 354,120 | | |
| 13 | 356,920 | | |
| 14 | 354,830 | | |

Mortality for the five-year period before the census was taken from the 2009 WHO life tables for Cambodia. The proportion surviving to exact age $x$, together with the associated logits for children aged less than 15 of both sexes and for women aged 10 to 65 are shown in Table 9.3.

*TABLE 9.3 WHO life table for Cambodia in 2009 used as standard*

| Children | $l(x)$ | Logit, $Y(x)$ |
|---|---|---|
| 0 | 1.0000 | |
| 1 | 0.9320 | –1.3085 |
| 5 | 0.9125 | –1.1723 |
| 10 | 0.9048 | –1.1256 |
| 15 | 0.8997 | –1.0968 |
| **Women** | | |
| 10 | 0.9138 | –1.1805 |
| 15 | 0.9099 | –1.1561 |
| 20 | 0.9049 | –1.1266 |
| 25 | 0.8980 | –1.0875 |
| 30 | 0.8889 | –1.0399 |
| 35 | 0.8760 | –0.9774 |
| 40 | 0.8603 | –0.9089 |
| 45 | 0.8419 | –0.8361 |
| 50 | 0.8179 | –0.7511 |
| 55 | 0.7847 | –0.6467 |
| 60 | 0.7374 | –0.5161 |
| 65 | 0.6732 | –0.3613 |

This life table was used as a standard against which to fit relational model life tables for children and adults separately for the periods 5–9 and 10–14 years before the census to the WHO life tables for Cambodia for 2000 and 1990. The fitted values of $\alpha$ and $\beta$ for children and adults are shown in Table 9.4.

*TABLE 9.4 Parameters of relational model life tables for five-year periods before the 2008 Census, using the 2009 WHO life table for Cambodia as the standard*

| | | Period before the census (years) | | |
|---|---|---|---|---|
| | | 0–4 | 5–9 | 10–14 |
| **Children** | $\alpha$ | 0.00 | 0.22 | 0.35 |
| | $\beta$ | 1.00 | 1.10 | 1.17 |
| **Adults** | $\alpha$ | 0.00 | 0.08 | 0.00 |
| | $\beta$ | 1.00 | 0.97 | 0.87 |

Estimates of age-specific fertility for mid-1997 and mid-2007 were obtained from the UN Population Division's

*World Population Prospects* (2010 revision) and are shown in Table 9.5.

*TABLE 9.5 Fertility schedules for Cambodia used in the reverse survival estimation*

| | Year | |
|---|---|---|
| Age group | 1997.5 | 2007.5 |
| 15–19 | 0.053 | 0.042 |
| 20–24 | 0.197 | 0.154 |
| 25–29 | 0.214 | 0.148 |
| 30–34 | 0.184 | 0.112 |
| 35–39 | 0.133 | 0.071 |
| 40–44 | 0.068 | 0.030 |
| 45–49 | 0.013 | 0.003 |

### *Step 1: Estimate the number of births in each year before the inquiry*

Applying Equations 3 and 4 to the logit values of the standard presented in Table 9.3, using the values of $\alpha$ and $\beta$ in Table 9.4, gives the estimates of $L_{x,T}$ and $P_{x,T}$ presented in Table 9.6.

Cohort survival to the census date of children aged 0 to 14 is then estimated using Equations 5 and 6 (presented in the last column of Table 9.6).

Finally, the number of births in each year preceding the census is derived by dividing the number of children enumerated in 2008 at each age by the relevant value of ${}^cL_x$ (Table 9.7).

### *Step 2: Estimate the mid-year populations of women by five-year age group*

Applying the values of $\alpha$ and $\beta$ for adult women in each five-year period given in Table 9.4 to the standard logits in Table 9.3 produces (by means of Equation 7) the five-year survivorship ratios for women by five-year age group shown in Table 9.8.

Application of the survivorship ratios from Table 9.8 to the enumerated population of women aged 10 to 64 using Equation 8 produces estimates of the population of women by five-year age group for 5, 10 and 15 years before the census (Table 9.9).

It is a straightforward arithmetic exercise to interpolate linearly between the four right-hand columns of Table 9.9 for each age group to produce estimates of the population by five-year age group for the 15 years centred on dates 6, 18, 30 … 174 months before the census, which is to say for 2007.67, 2006.67 … 1993.67. An extract of the results is shown in Table 9.10.

*TABLE 9.6 Calculation of cohort survival ratios by age, Cambodia 1983–2008*

| $x$ | $L_{x,0}$ | $L_{x,5}$ | $L_{x,10}$ | $P_{x,0}$ | $P_{x,5}$ | $P_{x,10}$ | ${}^cL_x$ |
|---|---|---|---|---|---|---|---|
| 0 | 0.9524 | 0.9439 | 0.9397 | 0.9524 | 0.9439 | 0.9397 | 0.9524 |
| 1 | 0.9298 | 0.9170 | 0.9106 | 0.9763 | 0.9716 | 0.9691 | 0.9298 |
| 2 | 0.9252 | 0.9112 | 0.9039 | 0.9951 | 0.9936 | 0.9926 | 0.9252 |
| 3 | 0.9197 | 0.9041 | 0.8958 | 0.9941 | 0.9923 | 0.9910 | 0.9181 |
| 4 | 0.9145 | 0.8974 | 0.8881 | 0.9944 | 0.9926 | 0.9914 | 0.9104 |
| 5 | 0.9118 | 0.8939 | 0.8840 | 0.9970 | 0.9960 | 0.9954 | 0.9049 |
| 6 | 0.9102 | 0.8919 | 0.8818 | 0.9983 | 0.9978 | 0.9974 | 0.9003 |
| 7 | 0.9087 | 0.8899 | 0.8795 | 0.9983 | 0.9978 | 0.9974 | 0.8954 |
| 8 | 0.9071 | 0.8879 | 0.8771 | 0.9983 | 0.9977 | 0.9973 | 0.8911 |
| 9 | 0.9056 | 0.8858 | 0.8747 | 0.9983 | 0.9977 | 0.9973 | 0.8871 |
| 10 | 0.9043 | 0.8842 | 0.8728 | 0.9986 | 0.9981 | 0.9978 | 0.8834 |
| 11 | 0.9033 | 0.8828 | 0.8713 | 0.9989 | 0.9985 | 0.9983 | 0.8801 |
| 12 | 0.9022 | 0.8815 | 0.8698 | 0.9989 | 0.9985 | 0.9983 | 0.8767 |
| 13 | 0.9012 | 0.8802 | 0.8682 | 0.9989 | 0.9985 | 0.9982 | 0.8741 |
| 14 | 0.9002 | 0.8789 | 0.8667 | 0.9989 | 0.9985 | 0.9982 | 0.8719 |

*TABLE 9.7 Estimated number of births in each of the 15 years preceding the 2008 Census, Cambodia*

| Mid-year preceding the census date | Births |
|---|---|
| 2007.67 | 295,328 |
| 2006.67 | 281,062 |
| 2005.67 | 290,118 |
| 2004.67 | 312,409 |
| 2003.67 | 306,453 |
| 2002.67 | 324,644 |
| 2001.67 | 325,997 |
| 2000.67 | 337,355 |
| 1999.67 | 354,583 |
| 1998.67 | 301,184 |
| 1997.67 | 370,118 |
| 1996.67 | 318,451 |
| 1995.67 | 403,924 |
| 1994.67 | 408,337 |
| 1993.67 | 406,976 |

*TABLE 9.8 Survival factors, ${}_5P_{x,T}$, by five-year age group and period before the Census, Cambodia 1993–1998*

| Age group | $T=0$ | $T=5$ | $T=10$ |
|---|---|---|---|
| 10–14 | 0.9951 | 0.9942 | 0.9945 |
| 15–19 | 0.9935 | 0.9922 | 0.9927 |
| 20–24 | 0.9912 | 0.9895 | 0.9902 |
| 25–29 | 0.9878 | 0.9855 | 0.9866 |
| 30–34 | 0.9838 | 0.9810 | 0.9826 |
| 35–39 | 0.9804 | 0.9772 | 0.9793 |
| 40–44 | 0.9752 | 0.9714 | 0.9744 |
| 45–49 | 0.9658 | 0.9609 | 0.9654 |
| 50–54 | 0.9501 | 0.9438 | |
| 55–59 | 0.9271 | | |

*TABLE 9.9 Number of women by five-year age group 0, 5, 10 and 15 years before the 2008 Cambodia Census*

| Age group | 2008.17 | 2003.17 | 1998.17 | 1993.17 |
|---|---|---|---|---|
| 10–14 | 815,930 | 784,134 | 705,852 | 640,512 |
| 15–19 | 780,320 | 701,746 | 636,977 | 372,754 |
| 20–24 | 697,160 | 632,014 | 370,014 | 454,008 |
| 25–29 | 626,430 | 366,133 | 449,548 | 414,964 |
| 30–34 | 361,650 | 443,047 | 409,395 | 376,470 |
| 35–39 | 435,880 | 401,624 | 369,909 | 320,300 |
| 40–44 | 393,760 | 361,474 | 313,672 | 258,772 |
| 45–49 | 352,520 | 304,701 | 252,137 | 190,119 |
| 50–54 | 294,280 | 242,285 | 183,546 | |
| 55–59 | 230,200 | 173,224 | | |
| 60–64 | 160,590 | | | |

*TABLE 9.10 Number of women by five-year age group by year*

| Age group | 2007.67 | 2006.67 | 2005.67 | 2004.67 | 2003.67 | 2002.67 |
|---|---|---|---|---|---|---|
| 10–14 | 812,750 | 806,391 | 800,032 | 793,672 | 787,313 | 776,305 |
| 15–19 | 772,463 | 756,748 | 741,033 | 725,318 | 709,604 | 695,269 |
| 20–24 | 690,645 | 677,616 | 664,587 | 651,558 | 638,528 | 605,814 |
| 25–29 | 600,400 | 548,341 | 496,281 | 444,222 | 392,162 | 374,474 |
| 30–34 | 369,790 | 386,069 | 402,349 | 418,628 | 434,907 | 439,682 |
| 35–39 | 432,454 | 425,603 | 418,752 | 411,901 | 405,050 | 398,453 |
| 40–44 | 390,531 | 384,074 | 377,617 | 371,160 | 364,703 | 356,694 |
| 45–49 | 347,738 | 338,174 | 328,610 | 319,046 | 309,483 | 299,444 |

***Step 3: Derive General Fertility Ratios***

General Fertility Ratios for each year are calculated using Equation 9, the births from Table 9.7 and the population of women aged 15 to 49 (from Table 9.10).

*TABLE 9.11 Estimates of the General Fertility Ratio, Cambodia, 1993–2008*

| Mid-year preceding the census date | GFR (15–49) |
|---|---|
| 2007.67 | 82 |
| 2006.67 | 80 |
| 2005.67 | 85 |
| 2004.67 | 93 |
| 2003.67 | 94 |
| 2002.67 | 102 |
| 2001.67 | 106 |
| 2000.67 | 112 |
| 1999.67 | 121 |
| 1998.67 | 106 |
| 1997.67 | 134 |
| 1996.67 | 119 |
| 1995.67 | 156 |
| 1994.67 | 163 |
| 1993.67 | 168 |

***Step 4: Estimate age-specific and Total Fertility***

Interpolating between the two fertility schedules in Table 9.5, fertility distributions – assuming Total Fertility of one child per woman – can be calculated for successive periods of a year preceding the census. An extract of the series, illustrating a portion of the calculations, is shown in Table 9.12.

For each year, multiplication of the number of women in

each age group by the corresponding fertility distribution for that age group produces the estimated number of births in that age group if Total Fertility was one child per woman.

However, the number of births each year has been estimated already (Table 9.7). The ratios of those estimates to the estimates based on the assumption that Total Fertility was 1 represent estimates of actual Total Fertility in each year.

**Interpretation**

The results of the application of reverse survival methods to data from the 2008 Census of Cambodia are plotted in Figure 9.1. The figure gives a strong visual sense of the progress of fertility decline in Cambodia over the 15-year period up to 2008. While the Crude Birth Rate estimated by Method 1 rose slightly between two five-year periods before the census, Total Fertility has continued to fall. Figure 9.1 also shows clearly the impact on the age structure, and thereby the estimates of fertility, of a preference for reporting children to be ages 8, 10 and 12, rather than 9 or 11.

In general, the main problems that are likely to arise with fertility estimates calculated by reverse survival methods stem not from the reverse survival procedure itself or from the assumptions made about the age pattern of fertility, but from age reporting errors or the failure of some respondents to report all young children who are household members. Omissions of young children and overestimation of their ages often produce a false impression that fertility decline has accelerated in the few years before the data were collected. No evidence exists that such reporting errors have occurred in Cambodia.

Because only a minority of children die during the 15 years after their birth, and in most contemporary populations quite a small minority of children at that, any errors in the assumptions made about children's mortality have a less than proportional impact on the estimates of fertility. Overestimates of child mortality produce overestimates of past fertility and vice versa. As the proportion of children that have died increases with age, the impact of such errors tends to be greater for the more distant estimates, resulting in a slight bias in the estimated trend in fertility. It is often easier to determine whether such biases exist in the data if it is possible to plot multiple series of fertility estimates on one chart rather than just examining a single set of results, as in this illustrative analysis.

## EXTENSIONS OF THE METHOD

A well-established extension of reverse survival methods for estimating fertility is the Own-Children Method. The

*TABLE 9.12 Interpolated fertility distributions by five-year age group and year*

| Age group | 2007.67 | 2006.67 | 2005.67 | 2004.67 | 2003.67 | 2002.67 |
|---|---|---|---|---|---|---|
| 10–14 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 15–19 | 0.0123 | 0.0126 | 0.0128 | 0.0131 | 0.0133 | 0.0136 |
| 20–24 | 0.0456 | 0.0465 | 0.0474 | 0.0483 | 0.0492 | 0.0502 |
| 25–29 | 0.0496 | 0.0500 | 0.0503 | 0.0506 | 0.0509 | 0.0513 |
| 30–34 | 0.0426 | 0.0423 | 0.0421 | 0.0418 | 0.0416 | 0.0413 |
| 35–39 | 0.0310 | 0.0304 | 0.0299 | 0.0293 | 0.0287 | 0.0282 |
| 40–44 | 0.0158 | 0.0153 | 0.0148 | 0.0143 | 0.0138 | 0.0133 |
| 45–49 | 0.0031 | 0.0029 | 0.0028 | 0.0026 | 0.0024 | 0.0022 |

*TABLE 9.13 Estimates of Total Fertility, Cambodia 1993–2007*

| Mid-year preceding the census date | Total Fertility |
|---|---|
| 2007.67 | 2.76 |
| 2006.67 | 2.69 |
| 2005.67 | 2.85 |
| 2004.67 | 3.15 |
| 2003.67 | 3.18 |
| 2002.67 | 3.46 |
| 2001.67 | 3.56 |
| 2000.67 | 3.78 |
| 1999.67 | 4.09 |
| 1998.67 | 3.58 |
| 1997.67 | 4.48 |
| 1996.67 | 3.89 |
| 1995.67 | 4.98 |
| 1994.67 | 5.07 |
| 1993.67 | 5.10 |

method has been documented comprehensively by Cho, Retherford and Cho (1986). The Own-Children Method also produces estimates of Total Fertility but, instead of using independent estimates of the age pattern of fertility to apportion births among mothers, seeks to link each child enumerated in an inquiry to his or her mother. This makes it possible to tabulate children by both their own age and the age of their mothers. These counts can then be reverse survived to calculate the number of children born each year in the past according to the age of their mothers then.

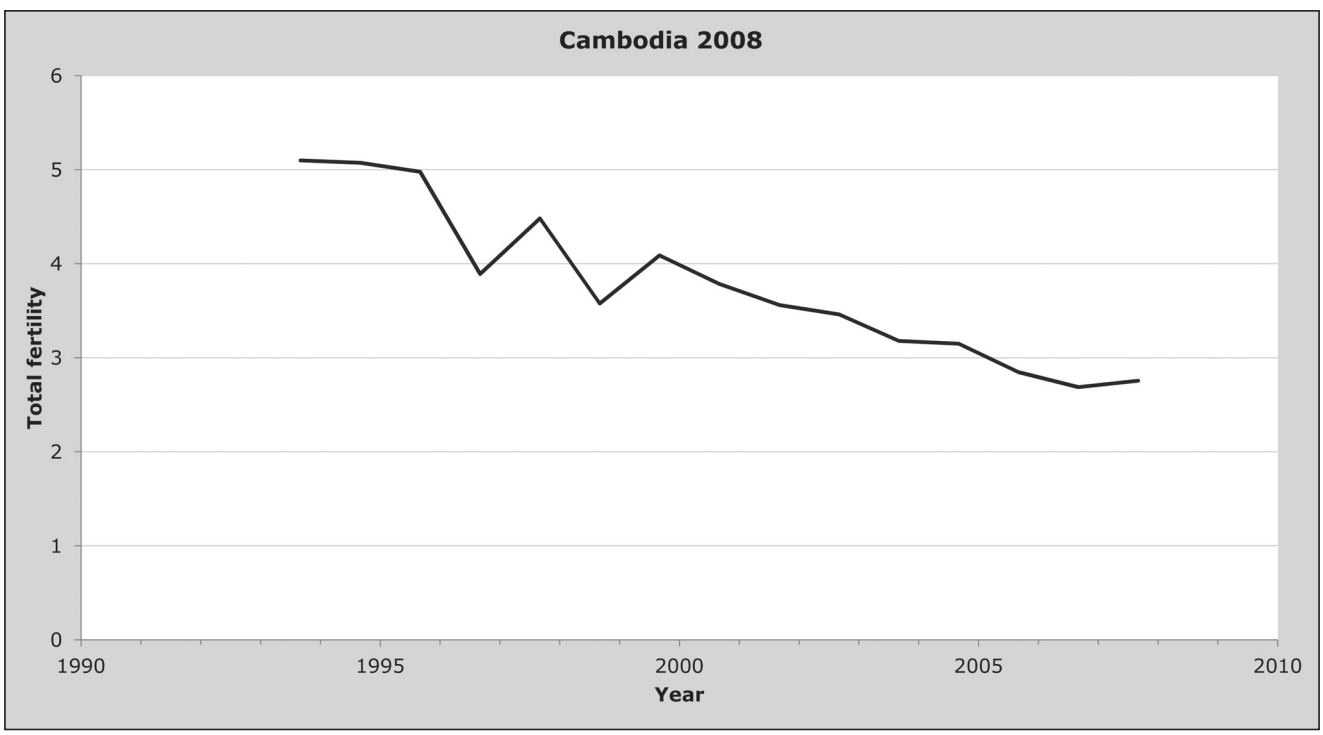
In inquiries in which detailed information has been collected on intra-household familial relationships and, in particular, if the identifier of co-resident mothers was included in the record for each child, this exercise in record linkage is somewhat algorithmically complex but do-able. Assumptions still need to be made about the age of the mothers of foster children and maternal orphans and, if they are numerous, it may be problematic to assume that they can be allocated to age groups of mothers on a *pro rata* basis.

If only limited information on relationships within the household is available, linking children to their mothers can become problematic, especially in populations where extended family households are common. In particular, if the mother is not resident in the household, children are at risk of being linked wrongly to an aunt, grandmother or other woman instead of remaining unlinked. The net impact of such errors is usually to inflate the estimated age-specific fertility of older women. Divorce, remarriage and polygynous marriage also increase the scope for identifying the wrong woman as the child's mother.

It is impossible to offer advice here on the detailed algorithms that should be used to link children to mothers when applying the Own-Children Method for estimating fertility, as the procedure adopted should be tailored to exactly what questions were asked and what fieldwork procedures were followed. For example, linkage becomes more straightforward if interviewers were instructed to list simple family units within a household together rather than, for example, grouping household members by age or gender.

Because substantial numbers of young people leave home in their late teenage years (including, in most cultures, young women who marry), the recommendation that

***FIGURE 9.1*** *Trend in Total Fertility estimated by reverse survival from the 2008 Cambodia Census*



reverse survival methods should be used only to estimate fertility for the 15 years before the inquiry concerned applies particularly strongly to estimates produced using the Own-Children Method.

Despite these potential limitations, experience suggests that the Own-Children Method can produce estimates of fertility of similar quality to those calculated from birth history data. Indeed, because they avoid the problem of truncation of the fertility rates for older women as the estimates move back in time and because birth history data must be collected from women personally and women who cannot be interviewed tend to have relatively lower fertility, it has been argued that estimates produced by the Own-Children Method are superior to direct estimates from birth histories in most instances. (Avery, St. Clair, Levin *et al*. 2013)

Equally, the reverse survival methods described in detail here usually perform well at a national level. They do not require access to individual-level data files and may produce more accurate estimates than the Own Children Method in inquiries in which matches of children to the wrong putative mother are common.

The Own-Children Method is particularly appropriate, however, for the study of differential fertility according to attributes of the mother such as her education, residence, or marital status. Because it can be applied to census data, the Own-Children Method can be used to estimate the fertility of sub-groups of the population that are too small to study using fertility survey data. Moreover, because far more information on women's attributes is usually collected in censuses than on birth certificates, the method remains of value even in countries with complete registration of births.

## REFERENCES


Avery C, T St. Clair, M Levin and K Hill. 2013. "The 'Own Children' fertility estimation procedure: A reappraisal", *Population Studies* **67**(2):171–183. doi: 10.1080/00324728.2013.769616

Brass W. 1979. "Evaluation of birth and death registration using age distributions and child survivorship data", *Asian and Pacific Census Forum* **5**(3):9–11,20. http://hdl.handle.net/10125/3563

Cho LJ, RD Retherford and MK Cho. 1986. *The Own-Child Method of Fertility Estimation*. Honolulu: University of Hawaii Press.

UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

# Multiple Census Methods

# Chapter 10 Synthetic relational Gompertz models

*Tom A Moultrie*

## DESCRIPTION OF METHOD

The synthetic relational Gompertz model is an extension of the relational Gompertz method for the estimation of age-specific and total fertility and makes use of two sets of parity data, collected at different points in time, together with estimates of current fertility for the intervening period based on reports of recent births classified by age.

The method explicitly allows changes in fertility to be taken into account and is designed to be applied to censuses or surveys conducted either 5 or 10 years apart. In such circumstances, the survivors of a cohort of women at the first inquiry can be identified at the second, and the change in the average parity of the cohort can be calculated. The resulting sequence of parity increments for different cohorts during the period between the inquiries can then be cumulated to calculate average parities for a hypothetical cohort experiencing the fertility implied by the observed parity increments.

The period fertility rates that are compared with these synthetic cohort estimates should ideally refer to the entire period between the two inquiries that asked about lifetime fertility. One way to ensure this is to make use of data on registered births classified by age of mother for each calendar year of the period. If such data are available, all births recorded during the period for each age group can be calculated by addition over calendar years. Average fertility rates for the period between the two inquiries can be obtained by dividing the births by the number of woman-years lived in each age group, estimated from the female population enumerated at the beginning and end of the period.

Where such data are not readily available, or are not reliable, a simpler, and generally adequate, procedure is to calculate age-specific fertility rates for the first and last years of the period, and to estimate the rates for the entire periods as the arithmetic mean of these two sets. If data on registered births are not available, but the two surveys or censuses gathered data on births in the past year, age-specific fertility rates for the period may by approximated in the same way by averaging the rates observed at the beginning and end of the period. If the births during the 12 months preceding each survey are tabulated by age of mother at the time of the survey, the observed fertility rates will correspond to age groups displaced by six months. The analysis will need to take this fact into account.

Once corresponding parities and fertility rates have been calculated for the period between the two inquiries, the cumulation and interpolation of the latter, and their comparison with the average parities, are carried out exactly as described in the presentation of the conventional relational Gompertz model.

## DATA REQUIRED

The data required are:

- The number of children ever born classified by five-year age group of mother, taken from two surveys or censuses five or 10 years apart.
- EITHER the number of births during the year preceding each survey classified by five-year age group of mother OR registered births by five-year age group of mother for each inter-survey year. If data on births classified by age of mother are not available for the end-points of the inter-survey period, an appropriate age-specific fertility schedule referring approximately to the middle of the period could be used.
- The number of women in each five-year age group from both surveys or censuses.
- If the crude birth rate is to be calculated, or the relative completeness of the data from the vital registration system is to be assessed, the total population recorded by each survey or census.

## ASSUMPTIONS

Most of the assumptions are those associated with the relational Gompertz model, namely:

- The standard fertility schedule chosen for use in the fitting procedure appropriately reflects the shape of the fertility distribution in the population.
- Any inter-survey changes in fertility have been smooth and gradual and have affected all age groups in a broadly similar way.
- Errors in the pre-adjustment fertility rates are proportionately the same for women in the central age groups (20–39), so that the age pattern of fertility described by reported births in the past year is reasonably accurate.
- The parities reported by younger women in their twenties are accurate.

The calculation of the synthetic cohort mean parities assumes that mortality and migration have no effect on actual parity distributions. In other words, it is assumed that the average parity of those women who die or migrate between the surveys is not significantly different from the average parity at comparable ages of those women who are alive and present at the end of the period.

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Before commencing analysis of fertility levels using this method, analysts should investigate the quality of the data at least in respect of the following dimensions:

- age and sex structure of the population;
- reported births in the last year; and
- average parities and whether an el-Badry correction is necessary.

## CAVEATS AND WARNINGS

It is crucially important that the sets of fertility rates being averaged are consistent with respect to age classification before they are averaged. If they are not consistent initially, because one refers to age groups displaced by six months and the other does not, the former set should be adjusted (for example, by applying the *F*-only variant of the relational Gompertz model) before proceeding. In general, estimates of age-specific fertility rates from different sources (e.g. vital registration and census) should not be combined because of the different ways in which the schedules may be distorted.

If age-specific fertility rates for the end-points of the period are not available, a set of rates referring approximately to the mid-point of the period could be used. It should be remembered that only the pattern of the inter-survey age-specific fertility rates is important in applying the relational Gompertz method, so that if this pattern was more or less constant over the period, the exact reference date of the rates used does not matter.

If data on registered births are used, changes in completeness of the data by age group over time could distort the pattern of fertility. If this has been the case, the method should be applied with caution.

## APPLICATION OF THE METHOD

The method is applied in the following steps.

### *Step 1: Calculation of reported average parities*

Calculate the average parities, ${}_5P_x(t_1)$ and ${}_5P_x(t_2)$ of women in each age group $[x,x+5)$ for the two inquiries ($t_1$ and $t_2$), for $x = 15,20\ldots\ 45$. For ease of exposition, we denote the average parity in age group $i$ at time $t$ by $P(i,t)={}_5P_x(t)$, where $i = (x/5–2)$. Thus, the average parities obtained from the first census or survey are denoted by $P(i,1)$, and those from the second survey by $P(i,2)$.

### *Step 2: Calculation of average parities for a hypothetical cohort*

The way in which the parities are calculated depends upon the length of the inter-survey interval.

#### *a) Interval is of five years' duration*

If the interval between the two data series is five years, all the survivors of age group $i$ at the first inquiry are in age group $i+1$ at the second inquiry, and the parity increment between the inquiries for the corresponding cohort is equal to $P(i+1,2)–P(i,1)$. Such increments can be calculated for each age group, and the hypothetical-cohort parities are then obtained by successively cumulating them. Thus, if the parity increment for the cohort of age group i at the first inquiry is denoted by $\Delta P(i+1)$, and the parity of age group $i$ for the hypothetical cohort is denoted by $P(i,s)$ (where the $s$ stands for 'synthetic'), one has $\Delta P(i+1) = P(i+1,2)–P(i,1)$ for $i = 1\ldots6$, and hence

$$P(i,s) = \sum_{j=1}^{i} \Delta P(j).$$

The parity increment $\Delta P(i+1)$ for the youngest age group ($i = 0$) is taken as being equal to $P(1,2)$, i.e., assuming that

$P(0,1)$ – the average parity of women aged 10–14 in the first inquiry– is zero. If fertility is changing rapidly, this value of $\Delta P(1)$ will therefore reflect period rates somewhat closer to the second survey than to the mid-point of the interval, slightly over-allowing for the change in fertility.

***b) Interval is of ten years' duration***

If the intercensal or inter-survey interval is 10 years, then the survivors of the initial cohort of age group $i$ in the first survey will be the women in age group $(i+2)$ in the second. The hypothetical cohort parities are then obtained by cumulating two parallel sequences of parity increments. Once more, for the youngest age groups, $\Delta P(1)$ is taken as being equal to $P(1,2)$ and $\Delta P(2)$ to $P(2,2)$. Other parity increments are calculated as $\Delta P(i+2) = P(i+2,2) - P(i,1)$ for $i = 1 \ldots 5$.

Hypothetical cohort parities for even-numbered age groups are obtained by summing the parity increments for even-numbered age groups, whereas those for odd-numbered age groups are obtained by summing parity increments for odd-numbered age groups. Thus,

$$
\begin{aligned}
P(1,s) &= \Delta P(1) = P(1,2) \\
P(2,s) &= \Delta P(2) = P(2,2) \\
P(3,s) &= \Delta P(1) + \Delta P(3) \\
P(4,s) &= \Delta P(2) + \Delta P(4) \\
P(5,s) &= \Delta P(1) + \Delta P(3) + \Delta P(5) \\
P(6,s) &= \Delta P(2) + \Delta P(4) + \Delta P(6) \\
P(7,s) &= \Delta P(1) + \Delta P(3) + \Delta P(5) + \Delta P(7) \, .
\end{aligned}
$$

***Step 3: Calculation of the current fertility rates***

The method of calculating this schedule, denoted by $f(i)$, where $i$ indexes the age groups as before, depends upon the data available.

***a) Data from a vital registration system***

One possible procedure is to calculate age-specific fertility rates referring roughly to the first and last years of the period between the two inquiries using data on the reported number of births during the year preceding each inquiry. In such a case, for each inquiry one would divide the reported births for each five-year age group of mother by the reported number of women in the same age group and then obtain age-specific fertility rates for the intervening period by calculating the arithmetic mean of each pair of end-point rates.

Alternatively, if age-specific fertility rates are available from a vital registration system for the whole period, a mean age-specific fertility rate for the period for each age group could be used. Calculating this mean would involve summing the births reported for each age group of mother, and dividing by the person years lived (by averaging the size of the age groups at the beginning and end of the interval, and multiplying by the number of years in the period).

Age-specific fertility rates obtained from vital registration are, by definition, classified by age of mother at the time of the delivery of the child.

***b) Data from the inquiries giving rise to the average parities in Step 2***

If the data on fertility are to be drawn from women's reports of recent fertility in the year before each of the surveys used to derive the average parities, the arithmetic mean of the two fertility schedules is still taken as the estimate of fertility in the intervening period. However, the schedule of fertility rates derived in this way applies to the six months before each survey, and hence the age classification of the rates must be adjusted to reflect the classification by age of mother at census, and not the birth of the child. This age shift in the rates must be taken into account in the application of the relational Gompertz model.

The process of fitting a relational Gompertz model to the data is exactly as described in Chapter 6. The only points of difference to note are the following:

- The estimates apply to the mid-point of the period, that is either 2½ or 5 years before the second inquiry.
- The spreadsheet only allows for the conventional application of the relational Gompertz model, using the parities to set the level and using the fertility schedule based on current fertility data for the intervening period to determine the shape of the fertility curve.
- If the data are classified by age of mother at the inquiry date (i.e. when the data on recent fertility are drawn from the census or survey that also provided the average parities, rather than from a vital registration system), the accompanying Excel workbook (see website) allows only for recent fertility data to be based on births reported in the 12 months preceding the census or survey.

The relevant steps are reproduced below.

***Step 4: Choose the fertility standard to be used with the model***

The default fertility standard is that produced by Booth, modified slightly by Zaba (1981). The standard is appropriate to high- and medium-fertility populations and is simply a normalized cumulated fertility schedule (i.e. with total fertility equal to one). The standard $Y^s(x)$ values, are determined by taking the gompits of the schedule. The standard parity values, $Y^s(i)$, are the gompits of the parities associated with the standard fertility schedule. The choice of standard determines the values of $g()$ and $e()$ used in the regression fitting procedures.

***Step 5: Evaluate the plot of P-points and F-points***

The plots of $z(x)–e(x)$ against $g(x)$, and $z(i)–e(i)$ against $g(i)$ on the same set of axes are then used as a diagnostic for identifying common errors and trends in the data, as discussed in the main article on the relational Gompertz model.

***Step 6: Fit the model by selecting the points to be used***

Initially, all points should be included in the model. The only exception is if the average parities in one age group are *higher* than the average parities in the next, in which case the gompit will be undefined and the model cannot be fitted using that point. (Such a situation cannot occur in a real cohort, but could arise in a synthetic cohort, either because of data error or during a time of rapidly changing fertility.)

If the parity and fertility data are internally consistent, the plots of $z()–e()$ against $g()$ should result in straight lines. Those *P*-points and *F*-points that cause each plot to deviate from a straight line should be excluded from the model. Ordinary linear regression (using least squares) is used to fit lines to the *P*-points and *F*-points, and to identify, sequentially, those points that do not fit neatly on a straight line. The intention is to seek the largest combination of *P*- and *F*-points that lie (almost) on the same line, and to use these to fit the model.

Points are selected for inclusion or exclusion using the following guidelines:

- A contiguous series of points must be included in the model. Sequentially, only the end-most points can be excluded. (The reason for this is that each point on the graph is the result of calculations involving the ratio of a pair of adjacent data values. If the analysis leads you to conclude that a data value is unreliable as a denominator of one of these ratios it is not logical to accept it as the numerator of the next ratio).
- *P*-points should be eliminated in preference to *F*-points. This is because the average parity data are generally more prone to age-specific errors than the fertility data.
- *P*-points which deviate clearly from the straight line based only on the other *P*-points, and *F*-points which deviate clearly from the straight line based only on the other *F*-points should be eliminated early on in the fitting process.
- *P*- and *F*-points at older ages should be eliminated in preference to those at younger ages since data at these ages are usually the least reliable and show the least consistency between lifetime and recent fertility. The exception to this relates to the data points for women under the age of 20. Small numbers of events, as is usual for these young women, frequently make the estimates of average parities or cumulated fertility unreliable.
- Where only a marginally worse fit is achieved with more points, this is to be preferred to a slightly better fit achieved with fewer points. The spreadsheet calculates the root mean squared error (RMSE),

$$RMSE = \sqrt{\frac{\sum\left((z()-e())-\left(\alpha+(\beta-1)^2\frac{c}{2}+\beta g()\right)\right)^2}{n}},$$

from the points used to fit the model. This statistic can assist with determining the optimal number of data points to which to fit if there is uncertainty as to which of two competing models is better. In this case, one should choose the model with the lower RMSE.

***Step 7: Assess the fitted parameters***

The values of $\alpha$ and $\beta$ that represent the best-fitting line joining the remaining *P*-points and *F*-points must be checked to confirm that they are not so far from their central values as to suggest that the standard chosen is inappropriate. A good fit is indicated if $-0.3<\alpha<0.3$, and if $0.8<\beta<1.25$.

If the parameters lie outside this range, one or both of the underlying data series are problematic or the standard is inappropriate. Experimentation with another standard or changing the selection of points should be done before proceeding further. If the parameters still lie outside the ranges above, the method should be regarded as inappropriate.

***Step 8: Fitted ASFRs and total fertility***

Having estimated the two parameters of the model, they can be applied to the standard values for the parities to obtain fitted values, $Y(i) = \alpha + \beta . Y^s(i)$. These are then converted back into measures of the cumulative proportion of fertility achieved by age group $i$ using the anti-gompit transformation. The anti-gompits based on the parity distributions indicate the proportion of fertility achieved by that age group. Dividing observed parity in each age group by these proportions produces a series of estimates of total fertility. Averaging these values across the sub-set of age groups that were used to estimate $\alpha$ and $\beta$ gives the fitted estimate of total fertility, $\hat{T}$.

Applying the same $\alpha$ and $\beta$ to the standard gompits for the ages that divide conventional age groups (i.e. 20, 25…50), applying the anti-gompit transformation, and multiplying by $\hat{T}$ produces a scaled cumulated fertility schedule. Differencing successive estimates of cumulated fertility and dividing by five produces the fitted fertility schedule for conventional age groups (15–19; 20–24 etc.) even if the data were initially classified with a half-year shift.

**WORKED EXAMPLE**

This example uses data collected in two Kenyan Censuses, a decade apart, in 1989 and 1999. Both censuses asked questions about births in the last year and lifetime fertility. The method has been implemented in an accompanying Excel workbook (see website).

***Step 1: Calculation of reported average parities***

An el-Badry correction was applied to the data from the 1989 Census – its application to Kenya is described here. By contrast, the data from the 1999 Census had evidently been edited prior to release, and no missing data were indicated. The average parities from the two censuses are shown in the first two columns of Table 10.1. From these data, it would appear that the lifetime fertility of older women has fallen by around 0.6 of a child over that decade. However, the increase in lifetime fertility among younger women is somewhat surprising.

***Step 2: Calculation of average parities for a hypothetical cohort***

The intercensal interval is 10 years (from 1989 to 1999). We therefore use the routine described in Step 2(b) on page 98 to derive the cohort average parities, shown in the last column of Table 10.1. As described at that step, $\Delta P(1) = P(1,2) = 0.2848$ and $\Delta P(2) = P(2,2) = 1.3640$, while $P(5,s) = \Delta P(1) + \Delta P(3) + \Delta P(5) = 0.2848 + (2.6073 - 0.2416) + (5.3867 - 3.2138) = 4.8234$.

*TABLE 10.1 Average parities by age group, Kenya, 1989 and 1999 Censuses*

| Age group | 1989 | 1999 | Hypothetical cohort parity $P(i,s)$ |
|---|---|---|---|
| 15–19 | 0.2416 | 0.2848 | 0.2848 |
| 20–24 | 1.5247 | 1.3640 | 1.3640 |
| 25–29 | 3.2138 | 2.6073 | 2.6505 |
| 30–34 | 4.7602 | 4.1432 | 3.9825 |
| 35–39 | 6.2390 | 5.3867 | 4.8234 |
| 40–44 | 7.1204 | 6.3818 | 5.6041 |
| 45–49 | 7.5103 | 6.9143 | 5.4987 |

It appears that omissions of children ever born may have occurred at older ages, as the hypothetical cohort parity at the oldest age group is somewhat lower than that of women in the hypothetical inter-survey cohort aged 40–44.

***Step 3: Calculation of current fertility rates***

The data available are women's reports of the month and year of their last birth in the year before each census. As described in Chapter 3, these reports can be converted into estimates of age-specific and total fertility by assuming that all births reported in the census month occurred before the census date, and pro-rating the births in the census month one year before the census. Doing so produces the direct estimates of age-specific and total fertility shown in Table 10.2. The last

*TABLE 10.2 Direct estimates of age-specific and total fertility, Kenya, 1989 and 1999 Censuses*

| Age group | 1989 | 1999 | Average fertility |
|---|---|---|---|
| 15–19 | 0.0679 | 0.1107 | 0.0893 |
| 20–24 | 0.2179 | 0.2381 | 0.2280 |
| 25–29 | 0.2309 | 0.2124 | 0.2217 |
| 30–34 | 0.1908 | 0.1728 | 0.1818 |
| 35–39 | 0.1458 | 0.1193 | 0.1326 |
| 40–44 | 0.0764 | 0.0583 | 0.0673 |
| 45–49 | 0.0351 | 0.0203 | 0.0277 |
| **Total Fertility** | **4.82** | **4.66** | **4.74** |

column, the estimate of inter-survey fertility, is derived by averaging the rates for 1989 and 1999 in each age group.

It is worth noting that the quality of reporting of fertility in the two censuses is poor. The levels of fertility implied by these data are substantially lower than those implied by the synthetic cohort parities, or from the value of total fertility of 5.3 children per woman obtained in the Demographic and Health Survey conducted in Kenya in 1993.

***Step 4: Choose the fertility standard to be used with the model***

The default fertility standard is that produced by Booth, modified slightly by Zaba (1981). No other peer-reviewed standard for female fertility exists.

***Step 5: Evaluate the plot of P-points and F-points***

We begin by fitting models using all the *P*- and *F*-points. The results are shown in the first plot on the *Diagnostic plots* sheet of the accompanying Excel workbook (see website).

***Step 6: Fit the model by selecting the points to be used***

Following the guidelines set out above, points are sequentially removed from the model to achieve a greater congruence of the *P*-points and the *F*-points. The best fit is found using the *P*-points for ages 20–39 and the *F*-points for ages 20–44 (Figure 10.1).

***Step 7: Assess the fitted parameters***

In this application, the fitted values of $\alpha$ (–0.0286) and $\beta$ (1.0042) lie comfortably within the set range.

***Step 8: Fitted ASFRs and total fertility***

The total fertility implied by the fitted model is 5.56 children per woman (Table 10.3), and applies, approximately, to August 1994, the model having accommodated the shift in the data arising from the classification of mother's age. This level of fertility is broadly consistent with the estimate of 5.3 children per woman from the 1993 Kenyan DHS, as well as with estimates arising from the application of the relational Gompertz method to each data set separately.

***FIGURE 10.1*** *Plot of z()–e() against g() after elimination of points, synthetic cohorts based on the 1989 and 1999 Kenyan Census data*

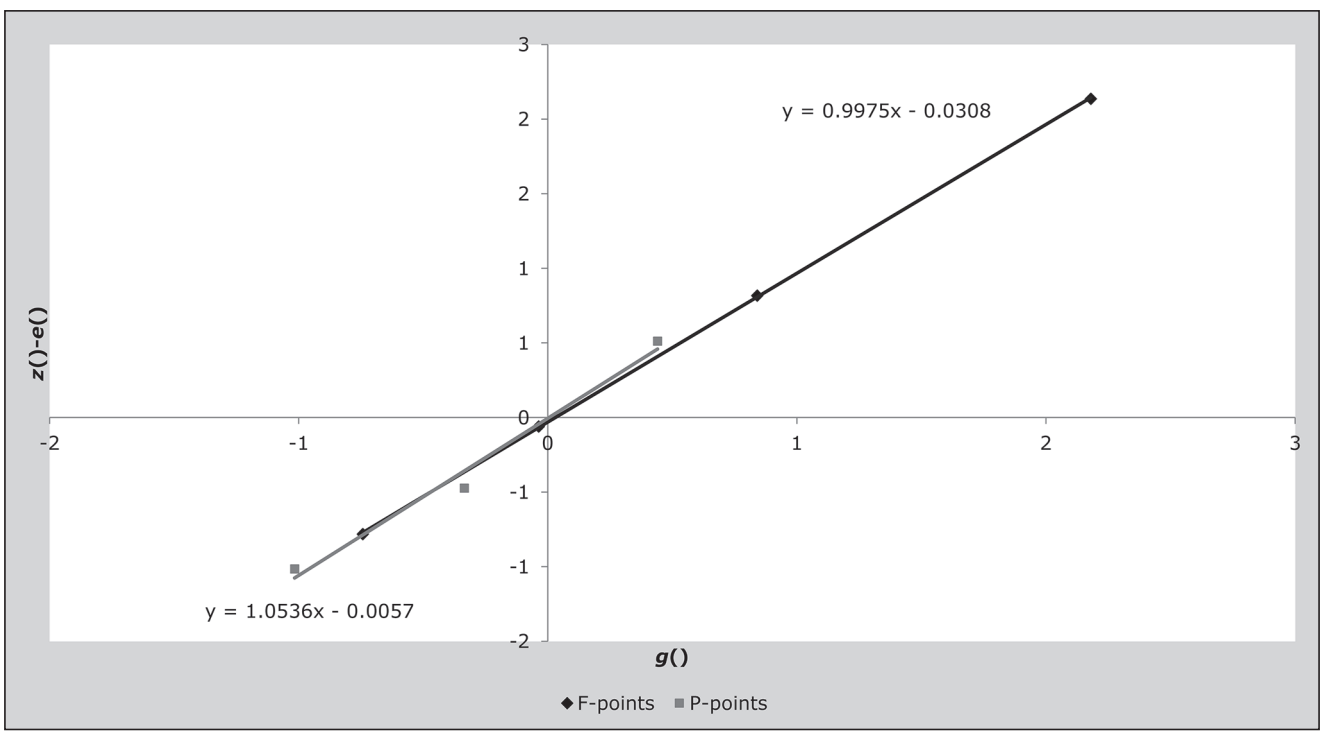


*TABLE 10.3 Estimated fertility rates based on hypothetical parity increments, Kenya 1989–1999*

| Age group | ASFR |
|---|---|
| 15–19 | 0.139 |
| 20–24 | 0.267 |
| 25–29 | 0.261 |
| 30–34 | 0.213 |
| 35–39 | 0.153 |
| 40–44 | 0.070 |
| 45–49 | 0.009 |
| **Total Fertility** | **5.56** |

## DETAILED DESCRIPTION OF METHOD

The method described here is – in effect – a variant of the relational Gompertz model that, instead of using parity and fertility data collected at one point in time, constructs an 'average' fertility schedule based on reports of current and lifetime fertility at two points in time. The mathematics of the relational Gompertz model is described fully in Chapter 7.

## FURTHER READING AND REFERENCES

This method was described initially by Zlotnik and Hill (1981) and re-presented on pages 41–45 of *Manual X* (UN Population Division 1983). The write-up here remains true to the original formulation, with the exception that it is presented as a variant of the relational Gompertz model where the parities used are the intercensal parities derived from the two surveys, and the fertility rates are the inter-survey estimates.


UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

Zaba B. 1981. *Use of the Relational Gompertz Model in Analysing Fertility Data Collected in Retrospective Surveys*. Centre for Population Studies Research Paper 81–2. London: Centre for Population Studies, London School of Hygiene & Tropical Medicine.

Zlotnik H and K Hill. 1981. "The use of hypothetical cohorts in estimating demographic parameters under conditions of changing fertility and mortality", *Demography* **18**(1):103–122. doi: 10.2307/2061052

# Chapter 11 Fertility estimates derived from cohort parity increments

*Tom A Moultrie*

## DESCRIPTION OF THE METHOD

Data on children ever born tabulated by standard five-year age group of women for a single census or survey convey a lot of information about the past fertility experience of the women. Unfortunately, however, if fertility has been changing, it is not possible to use the average parities of women in different age groups to obtain estimates of the age patterns of either cohort or period fertility.

If information on children ever born is available from two censuses or surveys approximately five or ten years apart, the change in the average number of children ever born by a particular cohort of women reflects their intercensal fertility. It then becomes possible to estimate an age-specific fertility schedule for the intervening period. Arretx (1973) developed a method for using such information with a 10-year interval between the inquiries. *Manual X* (UN Population Division 1983) presents a variant of an approach proposed by Coale and Trussell using the *P/F* ratio. A further refinement of the *Manual X* approach is presented here, based on the use of the relational Gompertz model.

The method estimates the average age-specific fertility rates in effect during the inter-survey period by constructing the average parities of a hypothetical, inter-survey cohort. A cumulated fertility schedule is then derived from these parities by interpolation, and age-specific fertility rates are obtained from cumulated fertility by successive subtraction.

The method is intended for situations in which it is possible to calculate average parities by age group of women for two points in time approximately five or ten years apart. If the interval between the inquiries is five years, the women in any five-year age group at the second inquiry represent the survivors of the women in the next younger five-year age group at the first inquiry. The difference in the average parity of the cohort between the first and the second inquiries reflects its childbearing experience between the two inquiries, if it is assumed that women who died or migrated between them had, on average, lifetime fertility that was not systematically different from that of the original women who remained. By cumulating the parity increments, it is possible to estimate average parities for a synthetic cohort experiencing throughout its hypothetical lifetime the age-specific fertility rates in effect during the period between the two inquiries. If the length of this period is ten years, a five-year age group at the second inquiry represents the survivors of the five-year age group who were two groups younger at the first inquiry. In this case, it is still possible to calculate the cohort parity increment for each cohort in order to construct the average parities of a hypothetical cohort. The method may be applied when the data come entirely or partially from nationally representative sample surveys as well as when they come from censuses, for although cohorts of particular individuals will not be identical on each occasion, their average parities will be representative of those of the sampled female population.

The two data sets need not refer to two points exactly five or ten years apart. For example, unless fertility is changing very rapidly, a four-year interval or an 11-year interval will provide reasonable estimates. In such a case, one is no longer following a cohort from survey to survey, but this factor is not very important because the average parity of an age group will not change rapidly from one year to the next.

Although the strength of method lies in its robustness to changing fertility, the technique presented here can also be used to estimate age-specific fertility rates using parity data from a single census or survey when fertility has not been changing during the reproductive life spans of the women concerned.

## ASSUMPTIONS

Most of the assumptions are those associated with the relational Gompertz model, namely

- The standard fertility schedule chosen for use in the fitting procedure appropriately reflects the shape of the fertility distribution in the population.
- Any changes in fertility have been smooth and gradual and have affected all age groups in a broadly similar way.
- The parities reported by younger women in their twenties are accurate.

Further, in deriving this measure of inter-survey fertility it is assumed that mortality and migration have no effect on actual parity distributions; that is, it is assumed that the average parity of those women who die or migrate between the surveys is not significantly different from the average parity at comparable ages of those women who are alive and present at the end of the period.

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Before commencing analysis of fertility levels using this method, analysts should investigate the quality of the data at least in respect of the following dimensions:

- age and sex structure of the population; and
- average parities and whether an el-Badry correction is necessary.

## CAVEATS AND WARNINGS

The general warning given about the use of information on children ever born in estimating fertility should be kept in mind in this instance. A tendency exists, even in countries with otherwise reasonably good data, for older women to omit some of their children, perhaps those who have died or those who have left home. As a result, average parities often fail to increase at a plausible rate, or may even decrease after age 35 or 40. The calculation of age-specific fertility rates from parities that suffer from such omissions will result in under-estimates of the fertility of older women. If the error is relatively minor, its effects may not be obvious. Thus, fertility estimates based on average parities of older women must be interpreted with caution, particularly if they indicate low fertility in relation to that estimated from the reports of younger women. Average parities for a hypothetical cohort are, moreover, very sensitive to changes in parity reporting from one inquiry to the other, and the calculation of such parities provides a useful consistency check of the raw data.

Whenever the additional data required on recent fertility exist, the procedure using a synthetic relational Gompertz model to compare cumulated intersurvey fertility rates with hypothetical-cohort average parities is to be preferred to the method described here, since the former method is less sensitive to the omission of children ever born from the reports of older women.

## APPLICATION OF THE METHOD

Steps 1 and 2 simply repeat the first two steps of the synthetic relational Gompertz method.

### *Step 1: Calculation of reported average parities from each inquiry*

Calculate the average parities, ${}_5P_x(t_1)$ and ${}_5P_x(t_2)$ of women in each age group $[x,x+5)$ for the two inquiries ($t_1$ and $t_2$), for $x = 15, 20\ldots 45$ if not already done as part of the preliminary investigations, or produced as a consequence of applying the el-Badry correction. For ease of exposition, we denote the average parity in age group $i$ at time $t$ by $P(i,t) = {}_5P_x(t)$, where $i = (x/5–2)$. Thus, the average parities obtained from the first census or survey are denoted by $P(i,1)$, and those from the second inquiry by $P(i,2)$.

### *Step 2: Calculation of average parities for a hypothetical cohort*

The way in which the parities are calculated depends upon the length of the interval between the two inquiries.

#### *a) Interval is of five years' duration*

If the interval between the two data series is five years, all the survivors of age group $i$ at the first inquiry are in age group $i+1$ at the second inquiry, and the parity increment between the inquiries for the corresponding cohort is equal to $P(i+1,2)–P(i,1)$. Such increments can be calculated for each age group, and the hypothetical-cohort parities are then obtained by successively cumulating them. Thus, if the parity increment for the cohort of age group $i$ at the first inquiry is denoted by $\Delta P(i+1)$, and the parity of age group $i$ for the hypothetical cohort is denoted by $P(i,s)$ (where the $s$ stands for 'synthetic'), one has $\Delta P(i+1) = P(i+1,2)–P(i,1)$ for $i = 1\ldots 6$, and hence

$$P(i,s) = \sum_{j=1}^{i} \Delta P(j).$$

The parity increment $\Delta P(i+1)$ for the youngest age group ($i = 0$) is taken as being equal to $P(1,2)$, i.e., assuming that

$P(0,1)$, the average parities of women aged 10–14 in the first inquiry, is zero. If fertility is changing rapidly, this value of $\Delta P(1)$ will therefore reflect period rates somewhat closer to the inquiry survey than to the mid-point of the interval, slightly over-allowing for the change in fertility.

***b) Interval is of ten years' duration***

If the intercensal or inter-survey period is 10 years, then the survivors of the initial cohort of age group $i$ in the first survey will be the women in age group $(i+2)$ in the second. Hypothetical cohort parities are then obtained by cumulating two parallel sequences of parity increments. Once more, for the youngest age groups, $\Delta P(1)$ is taken as being equal to $P(1,2)$ and $\Delta P(2)$ to $P(2,2)$. Other parity increments are calculated as $\Delta P(i+2) = P(i+2,2) - P(i,1)$ for $i = 1 \ldots 5$.

Hypothetical-cohort parities for even-numbered age groups are obtained by summing the parity increments for even-numbered age groups, whereas those for odd-numbered age groups are obtained by summing parity increments for odd-numbered age groups. Thus,

$$
\begin{aligned}
P(1,s) &= \Delta P(1) = P(1,2) \\
P(2,s) &= \Delta P(2) = P(2,2) \\
P(3,s) &= \Delta P(1) + \Delta P(3) \\
P(4,s) &= \Delta P(2) + \Delta P(4) \\
P(5,s) &= \Delta P(1) + \Delta P(3) + \Delta P(5) \\
P(6,s) &= \Delta P(2) + \Delta P(4) + \Delta P(6) \\
P(7,s) &= \Delta P(1) + \Delta P(3) + \Delta P(5) + \Delta P(7) \,.
\end{aligned}
$$

The following steps repeat those involved in using the relational Gompertz model, but fit a line only to the parity data.

***Step 3: Fitting of a relational Gompertz model***

If the parity data are internally consistent, the plots of $z(i) - e(i)$ against $g(i)$ should result in straight lines. Those $P$-points that cause the plot to deviate from a straight line should be excluded from the model. Ordinary linear regression (using least squares) is used to fit lines to the $P$-points and to identify, sequentially, those points that do not fit neatly on a straight line. The intention is to seek the most numerous combination of $P$-points that lie (almost) on the same line, and to use these to fit the model.

Points are selected for inclusion using the following guidelines:

- A contiguous series of points must be included in the model. Sequentially, only the end-most points can be excluded. (The reason for this is that each point on the graph is the result of calculations involving the ratio of a pair of adjacent data values. If the analysis leads to the conclusion that a data value is unreliable as a denominator of one of these ratios, it is not logical to accept it as the numerator of the next ratio.)
- $P$-points at older ages should be eliminated in preference to those at younger ages since data at these ages are usually the least reliable and exhibit the least consistency between lifetime and recent fertility.
- Where only a marginally worse fit is achieved with more points, this is to be preferred to a slightly better fit achieved with fewer points.

***Step 4: Assess the fitted parameters***

The values of $\alpha$ and $\beta$ that represent the best-fitting line joining the remaining $P$-points and $F$-points must be checked to ascertain that they are not so far from their central values as to suggest that the standard chosen is inappropriate. A good fit is indicated if $-0.3 < \alpha < 0.3$, and if $0.8 < \beta < 1.25$.

If the parameters lie outside this range, one or both of the underlying data series are problematic or the standard is inappropriate. Experimentation with another standard or changing the selection of points should be done before proceeding further. If the parameters still lie outside the ranges above, the method should be regarded as inappropriate.

***Step 5: Fitted ASFRs and total fertility***

Having estimated the two parameters of the model, they can be applied to the standard values for the parities to obtain fitted values, $Y(i) = \alpha + \beta . Y^s(i)$. These are then converted back into measures of the cumulative proportion of fertility achieved by age group $i$ using the anti-gompit transformation. The anti-gompits based on the parity distributions indicate the proportion of fertility achieved by that age group. Dividing observed parity in each age group by these proportions produces a series of estimates of total fertility. Averaging these values across the sub-set of age groups that were used to estimate $\alpha$ and $\beta$ gives the fitted estimate of total fertility, $\hat{T}$.

Applying the same $\alpha$ and $\beta$ to the standard gompits for the ages that divide conventional age groups (i.e. 20, 25…50), applying the anti-gompit transformation, and multiplying by $\hat{T}$ produces a scaled cumulated fertility schedule.

Differencing successive estimates of cumulated fertility and dividing by five produces the fitted fertility schedule for conventional age groups (15–19; 20–24 etc.).

These ASFRs are then deemed to apply to the mid-point of the period in between the two inquiries.

## WORKED EXAMPLE

The example uses the same data on average parities from the 1989 and 1999 Censuses of Kenya as in the example of the synthetic relational Gompertz model. In this application, however, it is assumed that the only available information is the average parities and that no data on recent fertility were collected. The process of fitting the relational Gompertz model to parity data alone is essentially similar to the basic relational Gompertz model. The exposition here therefore concentrates on the differences from that procedure. The method has been implemented in an accompanying Excel workbook (see website).

### *Step 1: Calculation of reported average parities from each inquiry*

An el-Badry correction was applied to the data from the 1989 Census. Its application is described here. By contrast, the data from the 1999 Census had evidently been edited, and no missing parity data were present. The average parities from the two censuses are shown in the first two columns of Table 11.1. From these data, it would appear that the cohort lifetime fertility of older women has fallen by around 0.6 of a child over that decade. However, the increase in lifetime fertility among teenaged women is somewhat surprising.

### *Step 2: Calculation of average parities for a hypothetical cohort*

The inter-survey period is 10 years (from 1989 to 1999). We therefore use the routine described in Step 2(b) on page 105 to derive the cohort average parities, shown in the last column of Table 11.1. As described in that step, $\Delta P(1) = P(1,2) = 0.2848$ and $\Delta P(2) = P(2,2) = 1.3640$, while $P(5,s) = \Delta P(1) + \Delta P(3) + \Delta P(5) = 0.2848 + (2.6073 - 0.2416) + (5.3867 - 3.2138) = 4.8234$. It is readily apparent that severe omissions of parities must have been present at older ages, as the hypothetical cohort parity at the oldest age group is somewhat lower than that of women in the hypothetical inter-survey cohort aged 40–44.

The definition of the age of the mother does not enter into this method. Average parities are – by definition – those prevailing at the survey or census date.

*TABLE 11.1 Average parities by age group, Kenya, 1989 and 1999 Censuses*

| Age group | 1989 | 1999 | Hypothetical cohort parity $P(i,s)$ |
|---|---|---|---|
| 15–19 | 0.2416 | 0.2848 | 0.2848 |
| 20–24 | 1.5247 | 1.3640 | 1.3640 |
| 25–29 | 3.2138 | 2.6073 | 2.6505 |
| 30–34 | 4.7602 | 4.1432 | 3.9825 |
| 35–39 | 6.2390 | 5.3867 | 4.8234 |
| 40–44 | 7.1204 | 6.3818 | 5.6041 |
| 45–49 | 7.5103 | 6.9143 | 5.4987 |

### *Step 3: Fitting of a relational Gompertz model*

The hypothetical cohort data in the last column of Table 11.1 are used to estimate fertility by means of the relational Gompertz model. Data points based on the average parities ($P$-points) are successively eliminated until the data points show a linear relationship with the (transformed) parities from the standard fertility schedule. The fitted points are shown in Figure 11.1.

Only five parity points can be plotted as the hypothetical parity for the 45–49 age group is lower than that of the 40–44 age group (5.4987 vs. 5.6041), meaning that the gompit of the ratio of this pair of points is undefined. Examining the points, there is evident under-reporting of fertility in the ages used to generate the last point plotted. Eliminating that point results in a much lower root mean square error, and the model is fitted to the remaining four points.

### *Step 4: Assess the fitted parameters*

The implied values of $\alpha$ and $\beta$ are –0.0084 and 1.0071 implying a fertility schedule fairly close to that underlying the modified Booth standard.

### *Step 5: Fitted ASFRs and total fertility*

Applying these parameters to the gompits of the parities in the standard using the linear relational model, $Y(i) = \alpha + \beta Y^s(i)$, taking the anti-gompits (column 4 of Table 11.2) and dividing these into the observed parities at the ages selected for inclusion in the model produces a series of five estimates of total fertility (ranging from 5.4 to 5.7 children per woman). Averaging these suggests total fertility ($\hat{T}$) is 5.54 children per woman.

*TABLE 11.2 Derivation of estimated total fertility (T-hat), Kenya, 1989 and 1999 Censuses*

| Age ($i$) | $Y^s(i)$ | Fitted $Y(i)$ | exp(–exp (–$Y(i)$)) | Actual cumulant |
|---|---|---|---|---|
| 0 | –2.0961 | –2.1194 | 0.0002 | 0.0013 |
| 1 | –1.0833 | –1.0994 | 0.0497 | 0.2754 |
| 2 | –0.3124 | –0.3230 | 0.2513 | 1.3930 |
| 3 | 0.3541 | 0.3482 | 0.4936 | 2.7368 |
| 4 | 1.0579 | 1.0570 | 0.7065 | 3.9166 |
| 5 | 1.9561 | 1.9615 | 0.8688 | 4.8167 |
| 6 | 3.4225 | 3.4384 | 0.9684 | 5.3688 |
| 7 | 6.0922 | 6.1270 | 0.9978 | 5.5320 |

*TABLE 11.3 Derivation of final adjusted fertility schedule, Kenya, 1989 and 1999 Censuses*

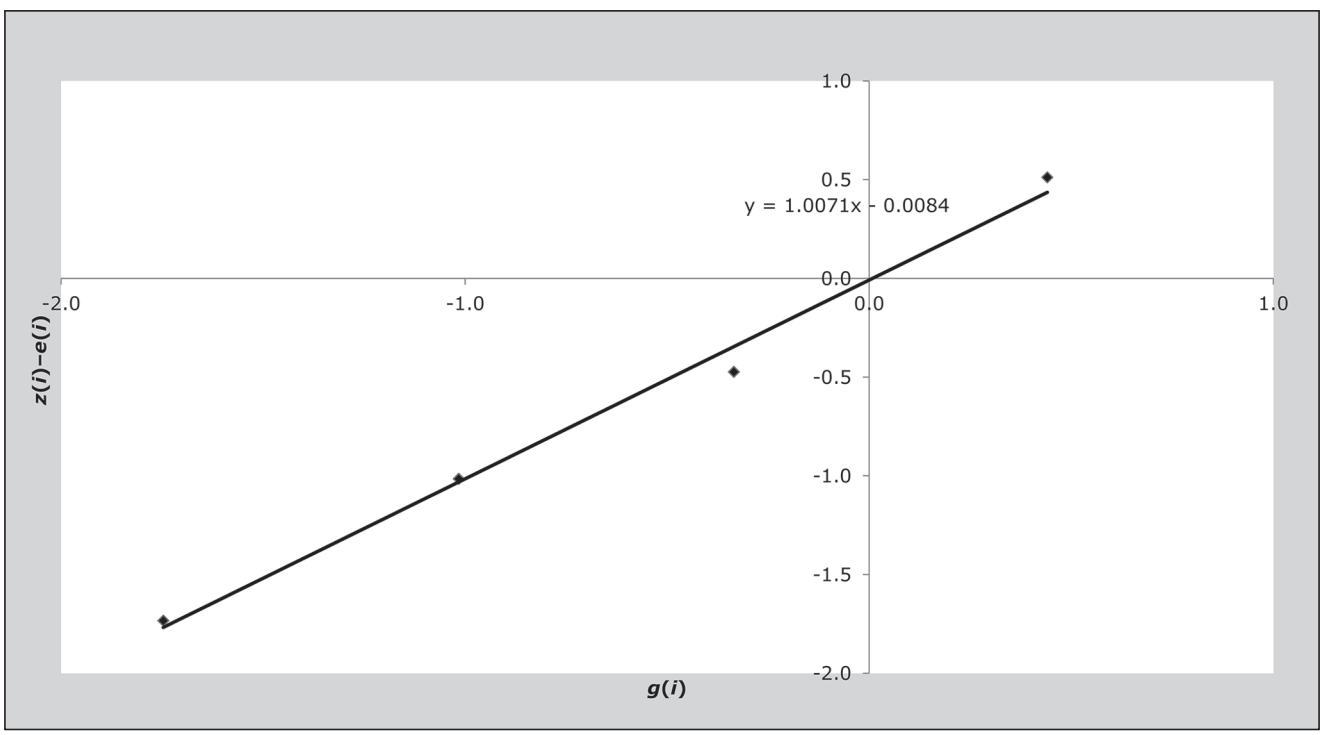
| Age ($x$) | $Y^s(x)$ | Fitted $Y(x)$ | exp(–exp (–$Y(i)$)) | $FM(x)$ | $fm(x)$ |
|---|---|---|---|---|---|
| 15 | –1.7731 | –1.7262 | 0.0036 | 0.0212 | 0.0042 |
| 20 | –0.6913 | –0.7318 | 0.1251 | 0.7318 | 0.1421 |
| 25 | 0.0256 | –0.0727 | 0.3411 | 1.9957 | 0.2528 |
| 30 | 0.7000 | 0.5472 | 0.5607 | 3.2801 | 0.2569 |
| 35 | 1.4787 | 1.2630 | 0.7537 | 4.4090 | 0.2258 |
| 40 | 2.6260 | 2.3176 | 0.9062 | 5.3013 | 0.1785 |
| 45 | 4.8097 | 4.3249 | 0.9869 | 5.7732 | 0.0944 |
| 50 | 13.8155 | 12.6034 | 1.0000 | 5.8501 | 0.0154 |
| **Total Fertility** | | | | | **5.53** |

Applying the fitted estimates of $\alpha$ and $\beta$ to the standard gompits, $Y^s(x)$, in each age group to derive the fitted gompits, then taking the anti-gompits and multiplying up by $\hat{T}$ produces the modified cumulative fertility schedule, $FM(x)$, below. Differencing and dividing by five produces the final schedule of age-specific fertility rates in the last column of Table 11.3.

The resulting estimate of total fertility is 5.5 children per woman, applicable half-way between the two censuses. In this application, the estimated age-specific fertility rates derived from the hypothetical-cohort parities can be compared with those obtained from the application of the synthetic relational Gompertz model (TFR = 5.56 children per woman). The similarity of the two sets of results is reassuring.

It must be remembered, however, that the results can be

*FIGURE 11.1 Fitted P-points, hypothetical cohorts, Kenya, 1989 and 1999 Census data*



seriously distorted if children ever born tend to be omitted from the reports provided by their mothers, particularly if the extent of such omission changes from one survey to the next.

## DETAILED DESCRIPTION OF METHOD

The method described here is simply a variant of the relational Gompertz model, but instead of using parity and fertility data collected at one point in time, constructs an 'average' fertility schedule based on reports of lifetime fertility at two points in time. It uses these alone to determine a fertility schedule. The mathematics of the relational Gompertz model is described fully in Chapter 7.

## VARIANTS OF THE METHOD

An option in the spreadsheet allows the intercensal period to be set to zero. This allows the derivation of TF from a single set of parity data. For this procedure to yield plausible estimates, not only would the average parities would have to be without error, but fertility would have had to have been constant for an extended period of time preceding the inquiry.

## FURTHER READING AND REFERENCES

The basic mechanics of the method were set out by Arretx (1973) and written up in *Manual X* (UN Population Division 1983). The version in *Manual X* used the *P/F* ratio method to convert the parity increments into fertility rates. The method described here uses the more versatile relational Gompertz method.


Arretx C. 1973. "Fertility estimates derived from information on children ever born using data from censuses," in *International Population Conference, Liège 1973*. Vol. 2. Liège: International Union for the Scientific Study of Population, pp. 247–261.

UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

# Survey Data Methods

# Chapter 12 Direct estimation of fertility from survey data containing birth histories

*Tom A Moultrie*

## DESCRIPTION OF METHOD

The direct estimation of fertility (age-specific, and total) from survey data containing birth histories is relatively straightforward. If the data are carefully collected with a validated instrument (such as that used by the Demographic and Health Surveys), they can provide reliable and accurate estimates of fertility. However, distortions also frequently occur in birth history data, especially in relation to the shifting of births to more distant years to avoid additional questions on, for example, child health or anthropometry (Cleland 1996). These problems have again been highlighted recently by Schoumaker (2010, 2011). Displacement and omission of births might cause fertility (particularly in the period three to five years before the survey) to be underestimated.

Two approaches can be used to estimate fertility directly from data containing a detailed birth history. The first approach – that used by the DHS in its official reports – produces an estimate covering the one- or three-year period before the survey. (Three-year estimates are frequently used to avoid undesirable fluctuations in the estimates arising from the relatively small number of annual births in the DHS). This approach is described in detail in the *Guide to DHS Statistics* (Rutstein and Rojas 2003). There are two disadvantages to it. First, if the survey is carried out over an extended period, it becomes impossible to locate the measure of fertility precisely in time. Second, the calculation of fertility rates is made more complex both by having to refer to the survey date and by working in five-year age groups and three-year periods of calendar time.

The simpler approach described here produces estimates of fertility for individual ages and calendar years of time. These can be very easily aggregated to produce estimates for wider age groups, or for periods of several years.

As with the DHS approach, initial manipulations have to be performed at a unit record level. For this reason, it makes sense in almost all circumstances to estimate fertility directly from birth histories using the built-in survival time functionality of a statistical analysis program such as *Stata*. A useful routine for performing these calculations in *Stata* has been produced by Schoumaker (2013). However, the calculations are sufficiently straightforward to carry out using simple cross-tabulations of data. This chapter describes how.

## DATA REQUIREMENTS AND ASSUMPTIONS

### Data required

Two sets of data, both routinely produced at the data processing stage of a survey with detailed birth histories, are required. The first is a data set in which the unit of analysis is the woman – i.e. there is one record per woman. These data are required to estimate the denominator of the fertility rates. The second data set has the child as the unit of analysis – i.e. there is one record per child – but also includes essential information on the mother (crucially, her date of birth) in each record in the data set.

To estimate fertility, the following information must be present in the data.

***A) Women's data set***

- The month and year of each woman's birth, derived if necessary from a century-month code (CMC).
- The month and year of interview.
- Any variables needed to adjust the data for the sampling design and sample weights.
- Important covariates by which one might wish to assess differentials in fertility, bearing in mind that covariates at the date of interview may not have applied at the time the events of interest (recent births) took place.

***B) Child's data set***

- The child's date of birth – month and year, derived if necessary from a CMC.
- The mother's date of birth – month and year, derived if necessary from a CMC.
- Any variables needed to adjust the data for the sampling design and sample weights.
- The same covariates by which differentials in fertility are to be assessed.

## CAVEATS AND WARNINGS

- While single-age fertility rates derived from relatively small-scale surveys provide some indication of the quality of the data, the rates are almost always too erratic to be of direct use. Aggregation into five-year groups (and then – perhaps – smoothing the rates by means of a relational Gompertz model) is almost always called for.
- Similarly, rates for a single calendar year derived from survey data may not be reliable. Data for multiple calendar years should be combined to produce a more reliable estimate. However, ideally, one should not combine more than three years' data to avoid flattening out the trend in fertility.
- The rates produced using this approach may be affected materially by omission or displacement of the date of reported births.
- The rates produced in this manner will not be the same as those produced by MeasureDHS. In the first place, the estimation of the period exposed to risk is a little different (MeasureDHS works in complete months, while here we work in half-months). Second, the reference period for the rates may differ by up to 11 months. One could, however, calculate rates for years running from July to June (and thus centred on 1 January, or indeed for any other 12-month period) by manipulating the numerator and denominator appropriately.

## APPLICATION OF METHOD

We define the following terms:

- $M_B^c$ – the child's month of birth
- $Y_B^c$ – the child's year of birth
- $M_B^m$ – the mother's month of birth
- $Y_B^m$ – the mother's year of birth
- $M_1$ – the month in which the mother is interviewed
- $Y_1$ – the year in which the mother is interviewed
- $B(x,t)$ – the total number of births to mothers aged $x$ at the birth of their child in calendar year $t$
- $E(x,t)$ – the person-years of exposure to risk of women aged $x$ in calendar year $t$.

The rates are calculated by means of the following steps. To avoid having to make additional assumptions about the exposure to risk in the month of interview, both exposure and births occurring in the month of interview are ignored.

The general case is presented below where not all women are interviewed in the same calendar year. Where all women are interviewed in the same calendar year, the process can be simplified accordingly.

***Step 1: Produce a tabulation of the number of births in each calendar year by the age of the mother at the birth of the child***

This step produces the numerator of the fertility rates: births of children by calendar year and age of mother at birth.

In principle, the tabulation is relatively straightforward, although care needs to be taken to allocate appropriately mother's age at the birth of her child when both mother and child have the same month of birth. If, as is usually the case, information on day of birth is not available, it is necessary to allocate the mother's day of birth randomly to fall before or after the child's day of birth. This could be implemented by generating a binary variable, $b$, using a random number generator, but doing so would have implications for the consistency and replicability of investigations. Instead, $b$ can be generated from a putatively uniform variable that has no bearing on the outcomes being investigated, such as the day of the month in which the mother was interviewed. We therefore define $b = 1$ if the day of interview is greater than 15, and 0 if the day of the month is 15 or less.

The age (at last birthday) of the mother at the birth of a given child, $x$, is given by

$$x = \text{int}\left(\frac{12(Y_B^c - Y_B^m) + (M_B^c - M_B^m - b)}{12}\right)$$

where int() represents the integer portion of the term in brackets.

Extract a tabulation showing the total number of births in each cell defined by combinations of $Y_B^c$ and $x$, $B(x,t)$, weighting the data as appropriate, and making sure to exclude births that occurred in the month that the mother was interviewed.

***Step 2: Calculate the age of each woman at the start of the year in which she was interviewed***

Working with the women's data set (i.e. with one record per woman), begin by deriving the age of women on 1 January of the year of interview, $x_I$, assuming that mothers' births are uniformly distributed over calendar months (and hence occur, on average, half-way through each month):

$$x_I = Y_I^m - Y_B^m - 1. \quad (1)$$

It follows that the age of the mother on 1 January of any other year, $t$, ($t \le Y_I$) will be $x_I - (Y_I - t)$.

***Step 3: Calculate the exposed to risk for each woman in the year of her interview***

In the calendar year in which she is interviewed, a woman is exposed to the risk of giving birth for only a portion of the year (that is, the portion before the interview takes place). In this case, the computation of exposure to risk depends on whether the interview took place before or after the woman's birthday in that year. If her birth month precedes the interview month, she will be exposed to risk of giving birth at age $x_I$ for $E(x_I, Y_I) = \frac{M_B^m - 0.5}{12}$ years, and for $E(x_I + 1, Y_I) = \frac{M_I - M_B^m - 0.5}{12}$ years at age $x_I + 1$.

In contrast, if her birth month is the same as, or after, the month of her interview, her exposure to risk of giving birth in the year of interview will be for $E(x_I, Y_I) = \frac{M_I - 1}{12}$ years at age $x_I$, and $E(x_I + 1, Y_I) = 0$ years at age $x_I + 1$.

Note that in the last complete year, aggregate exposure per woman is 1 year, whereas in the year of interview, aggregate exposure is $(M_I - 1)/12$ of a year, regardless of the relative timing of birth month and interview month.

Variables giving each woman's exposure at ages $x_I$ and $x_I + 1$ in the year of interview must be derived, and then aggregated (weighting were necessary) to produce a tabulation of aggregate exposure by age in the year of interview.

***Step 4a: Calculate the exposure to risk for each woman in the last complete calendar year before her interview***

In the last complete calendar year before each woman is interviewed, i.e. in year $t = Y_I - 1$, she will be aged $x_I - 1$ until her birthday, and $x_I$ for the remainder of the year. On the same assumption as above of a uniform distribution of births within calendar months, the fraction of a year from 1 January until each woman's birthday is given by

$$E(x_I - 1, Y_I - 1) = \frac{M_B^m - 0.5}{12}$$

while for the remaining fraction of the year, she will be aged $x_I$ with exposure

$$E(x_I, Y_I - 1) = 1 - E(x_I - 1, Y_I - 1) = 1 - \frac{M_B^m - 0.5}{12}.$$

Using the two formulae above, variables giving each woman's exposure at ages $x_I$ and $x_I + 1$ in year $Y_I - 1$ must be derived, and then aggregated (weighting were necessary) to produce a tabulation of aggregate exposure by age in that year.

***Step 4b: Derive the exposure for earlier complete calendar years***

Birth histories are collected retrospectively from all women and each woman provides information for the entire period over which she has been exposed to the risk of childbearing. Some women may have moved between places or changed their other characteristics at some point during this period but, because complete residential and economic histories are seldom collected in fertility surveys, it is usually impossible to allow for this when calculating fertility rates. This means that the interpretation of some results such as fertility by place of residence becomes less clear.

However, since birthdays are immutable, and the population of women being assessed is constant over time, the aggregate exposure of women attaining age $x$ in a year for which all women's exposure is complete, $v$, will also equal the exposure of the cohort in earlier years, that is:

$$\begin{aligned} E(x, v-1) &= E(x-1, v-2) \\ &= \ldots = E(x-k, v-k-1). \end{aligned} \quad (2)$$

***Step 5: Derive the age-specific fertility rates***

The total exposure at each age in each calendar year, $E(x,t)$, is derived by summing the tabulations derived in Steps 3 and 4 for each age and for each calendar year (complete and incomplete). Note that if fieldwork extends over two calendar years, $Y_I - 1$ will refer to two different years, as will $Y_I$. Total exposure in the final calendar year for which exposure might be derived will be based on only the partial exposure of women interviewed in the final calendar year

of fieldwork, whereas total exposure in the immediately preceding year will be comprised of the partial exposure of women interviewed in the first year of fieldwork and the full exposure in that year of women interviewed in the final year of fieldwork.

The age-specific fertility rates for age $x$ in year $t$ are given by

$$f_x(t) = \frac{B_x(t)}{E_x(t)}.$$

Age-specific fertility rates for conventional five-year age groups are derived by summing the births to women across each age group, and dividing by the sum of the exposure in that age group. Thus, if $i = (x/5)-2$ for $x = 15,20,\ldots, 45$, then

$$f(1) =_5 f_{15}; f(2) =_5 f_{20}; \ldots f(7) =_5 f_{45}$$

and

$$f(i,t) = \frac{\sum_{a=5i+10}^{5i+14} B_a(t)}{\sum_{a=5i+10}^{5i+14} E_a(t)}.$$

To combine data for multiple years, the numerators and denominators are summed separately before dividing to produce the rate:

$$f\left(i,(t_1,t_2)\right) = \frac{\sum_{z=t_1}^{t_2}\sum_{a=5i+10}^{5i+14} B_a(z)}{\sum_{z=t_1}^{t_2}\sum_{a=5i+10}^{5i+14} E_a(z)}.$$

## WORKED EXAMPLE

This example uses data from the 2004 Malawi DHS. Fieldwork in this survey began in earnest in October 2004 and ran through to February 2005.

### *Step 1: Produce a tabulation of the number of births in each calendar year by the age of the mother at the birth of the child*

After random allocation of mother's age at birth in cases where the mother and child's month of birth are the same, the full cross-tabulation of children's year of birth by age of mother at the birth of her child is shown in Table 12.1. It would appear that there has been extreme shifting or omission of births in 2001 and 2002 in that the number of births reported in those years is some 20 per cent lower than that reported in 2003. Reported births in 2004 are lower than in 2003 in part because many women were not exposed for the full calendar year, and because births occurring in the month of interview are excluded from the analysis.

*TABLE 12.1 Classification of births since 2001 by age of mother at birth, Malawi, 2004 DHS*

| | Year of birth | | | | |
|---|---|---|---|---|---|
| **Age** | **2001** | **2002** | **2003** | **2004** | **2005** |
| 13 | 1.11 | 0.96 | 0.00 | 0.00 | 0.00 |
| 14 | 6.44 | 3.26 | 2.00 | 4.02 | 0.00 |
| 15 | 19.70 | 12.74 | 17.21 | 14.65 | 0.00 |
| 16 | 49.84 | 41.40 | 49.87 | 39.00 | 0.00 |
| 17 | 93.45 | 88.79 | 93.36 | 61.67 | 0.00 |
| 18 | 113.79 | 133.70 | 153.38 | 110.40 | 0.00 |
| 19 | 145.63 | 148.18 | 162.51 | 162.48 | 0.00 |
| 20 | 146.03 | 166.63 | 177.72 | 155.24 | 0.00 |
| 21 | 159.60 | 137.76 | 179.68 | 174.46 | 0.00 |
| 22 | 137.50 | 128.60 | 147.12 | 148.44 | 0.00 |
| 23 | 115.15 | 110.30 | 173.94 | 138.36 | 2.12 |
| 24 | 109.24 | 96.07 | 144.74 | 149.19 | 0.00 |
| 25 | 113.58 | 93.61 | 105.37 | 117.68 | 0.00 |
| 26 | 82.08 | 69.68 | 107.11 | 105.36 | 0.00 |
| 27 | 74.37 | 77.16 | 129.50 | 105.48 | 0.00 |
| 28 | 66.31 | 66.14 | 73.87 | 91.96 | 0.00 |
| 29 | 62.92 | 63.28 | 75.42 | 80.13 | 0.00 |
| 30 | 55.93 | 55.44 | 76.98 | 68.16 | 0.00 |
| 31 | 55.89 | 42.38 | 59.05 | 56.76 | 0.00 |
| 32 | 55.11 | 72.47 | 59.85 | 61.36 | 0.00 |
| 33 | 34.74 | 54.08 | 72.14 | 41.23 | 0.00 |
| 34 | 28.09 | 44.41 | 67.04 | 52.00 | 0.00 |
| 35 | 50.00 | 25.28 | 41.26 | 48.16 | 0.00 |
| 36 | 41.61 | 33.88 | 27.42 | 33.56 | 0.00 |
| 37 | 30.57 | 25.46 | 48.50 | 30.46 | 0.00 |
| 38 | 24.47 | 32.07 | 31.55 | 36.85 | 0.00 |
| 39 | 23.05 | 16.87 | 39.64 | 22.38 | 0.00 |
| 40 | 16.95 | 20.66 | 12.56 | 26.47 | 0.00 |
| 41 | 19.67 | 9.72 | 17.17 | 9.87 | 0.00 |
| 42 | 12.44 | 7.72 | 9.79 | 8.89 | 0.00 |
| 43 | 9.43 | 10.35 | 17.32 | 9.15 | 0.00 |
| 44 | 4.17 | 10.98 | 7.11 | 11.11 | 0.00 |
| 45 | 4.94 | 4.86 | 3.63 | 4.29 | 0.00 |
| 46 | 4.02 | 9.07 | 14.65 | 4.96 | 0.00 |
| 47 | 0.00 | 0.82 | 3.96 | 2.35 | 0.00 |
| 48 | 0.00 | 0.00 | 2.16 | 0.00 | 0.00 |
| 49 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **TOTAL** | **1967.84** | **1914.75** | **2404.58** | **2186.55** | **2.12** |

***Step 2: Calculate the age of each woman at the start of the year in which she is interviewed***

The age of women at the start of the year in which she is interviewed is derived from Equation 1. A sample extract is shown in Table 12.2. In the third line, the woman (case id 444 3) was born in August 1984 and interviewed in October 2004. On 1 January 2004 she would have been aged 19 (column 4). The woman with case id 528 2, in the ninth (penultimate) line of data, born in January 1970, interviewed in January 2005, and would have been aged 34 on 1 January 2005.

***Step 3: Calculate the exposure to risk for each woman in the year of her interview***

Columns (5) and (6) of Table 12.2 show the derivation of the exposure to risk for each woman in the year of her interview. The woman in the first line (case id 443 4) had her 28th birthday in February 2004. On the assumption that birthdays occur, on average, half-way through each month, she would have spent 0.125 (1.5 /12) aged 27 in 2004, and a further 0.625 of a year (7.5 months from the middle of February to the end of September, the month before she was interviewed) aged 28 in 2004.

The woman in the second line (case id 443 10) had her birthday in the same month she was interviewed. As a result, she experiences a full 9 months (0.75 of a year) exposure aged 29 in 2004, and has no exposure thereafter.

All women interviewed in January 2005 have no exposure in the year of interview, as we do not consider exposure (or births) that occur in that month.

***Step 4a: Calculate the exposure to risk for each woman in the last complete calendar year before her interview***

Columns (7) and (8) of Table 12.2 show the derivation of exposure to risk in the last complete year for which women were exposed to risk of giving birth in the survey data. For women interviewed in 2004, this would have been in 2003. For women interviewed in 2005, this would have been in 2004.

In the second case (case id 443 10), exposure in 2003 – her last complete year of exposure – would have been 9.5 months at age 28 and 2.5 months at age 29. As suggested by Equation 2, in previous years her exposure would have been distributed similarly, at commensurately younger ages: in 2002, exposure would have been 9.5 months at age 27 and 2.5 months at age 28.

In the last case presented (case id 529 2), the woman would have spent approximately 9.5 months (0.792 of a year) aged 31 in 2004, and 2.5 months (0.208 of a year) aged 32 in 2004.

Aggregating exposure by single year of age and calendar year from Step 4 produces the exposure to risk shown in Table 12.3.

*TABLE 12.2 Data showing derivation of exposure to risk, Malawi, 2004 DHS*

| | | | | Exposure in year of interview | | Exposure in last complete year | |
|---|---|---|---|---|---|---|---|
| caseid | Date of birth | Date of interview | Age at start of year of interview | Lower age | Higher age | Lower age | Higher age |
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| 443 4 | February 1976 | October 2004 | 27 | 0.125 | 0.625 | 0.125 | 0.875 |
| 443 10 | October 1974 | October 2004 | 29 | 0.750 | 0.000 | 0.792 | 0.208 |
| 444 3 | August 1984 | October 2004 | 19 | 0.625 | 0.125 | 0.625 | 0.375 |
| 445 2 | June 1983 | October 2004 | 20 | 0.458 | 0.292 | 0.458 | 0.542 |
| 519 7 | May 1989 | January 2005 | 15 | 0.000 | 0.000 | 0.375 | 0.625 |
| 522 2 | March 1979 | January 2005 | 25 | 0.000 | 0.000 | 0.208 | 0.792 |
| 526 4 | December 1989 | January 2005 | 15 | 0.000 | 0.000 | 0.958 | 0.042 |
| 526 7 | September 1979 | January 2005 | 25 | 0.000 | 0.000 | 0.708 | 0.292 |
| 528 2 | January 1970 | January 2005 | 34 | 0.000 | 0.000 | 0.042 | 0.958 |
| 529 2 | October 1972 | January 2005 | 32 | 0.000 | 0.000 | 0.792 | 0.208 |

*TABLE 12.3 Aggregate exposure by single year of age and calendar year, Malawi, 2004 DHS*

| Age | 2002 | 2003 | 2004 | 2005 |
|---|---|---|---|---|
| 11 | 0.063 | 0.000 | 0.000 | 0.000 |
| 12 | 198.291 | 0.063 | 0.000 | 0.000 |
| 13 | 468.833 | 198.291 | 0.063 | 0.000 |
| 14 | 432.083 | 468.833 | 197.506 | 0.000 |
| 15 | 490.890 | 432.083 | 409.831 | 0.049 |
| 16 | 522.245 | 490.890 | 370.078 | 0.402 |
| 17 | 597.259 | 522.245 | 431.191 | 0.216 |
| 18 | 606.502 | 597.259 | 444.050 | 0.337 |
| 19 | 594.975 | 606.502 | 528.989 | 0.622 |
| 20 | 573.166 | 594.975 | 514.654 | 0.674 |
| 21 | 480.330 | 573.166 | 521.777 | 0.354 |
| 22 | 574.521 | 480.330 | 489.303 | 1.172 |
| 23 | 486.871 | 574.521 | 422.082 | 0.166 |
| 24 | 405.933 | 486.871 | 503.468 | 0.939 |
| 25 | 405.592 | 405.933 | 416.489 | 0.729 |
| 26 | 407.569 | 405.592 | 350.520 | 0.000 |
| 27 | 346.264 | 407.569 | 354.229 | 0.425 |
| 28 | 313.426 | 346.264 | 349.949 | 0.265 |
| 29 | 286.749 | 313.426 | 300.703 | 0.337 |
| 30 | 308.209 | 286.749 | 262.300 | 0.177 |
| 31 | 252.422 | 308.209 | 252.010 | 0.000 |
| 32 | 309.337 | 252.422 | 256.686 | 0.166 |
| 33 | 267.239 | 309.337 | 217.728 | 0.000 |
| 34 | 183.176 | 267.239 | 271.954 | 0.000 |
| 35 | 185.172 | 183.176 | 226.209 | 0.868 |
| 36 | 222.879 | 185.172 | 151.012 | 0.000 |
| 37 | 217.592 | 222.879 | 166.838 | 0.000 |
| 38 | 236.389 | 217.592 | 192.603 | 0.110 |
| 39 | 177.195 | 236.389 | 194.856 | 0.363 |
| 40 | 161.461 | 177.195 | 195.769 | 0.591 |
| 41 | 142.134 | 161.461 | 155.461 | 0.000 |
| 42 | 173.338 | 142.134 | 133.356 | 0.166 |
| 43 | 168.616 | 173.338 | 126.403 | 0.000 |
| 44 | 148.788 | 168.616 | 147.170 | 0.088 |
| 45 | 140.768 | 148.788 | 143.087 | 0.088 |
| 46 | 138.297 | 140.768 | 125.995 | 0.000 |
| 47 | 72.711 | 138.297 | 124.497 | 0.000 |
| 48 | 0.606 | 72.711 | 117.910 | 1.027 |
| 49 | 0.000 | 0.606 | 53.140 | 0.000 |
| **TOTAL** | **11697.89** | **11697.89** | **10119.87** | **10.330** |

*TABLE 12.4 Age-specific fertility rates by single years of age and calendar year, Malawi, 2004 DHS*

| Age | 2001 | 2002 | 2003 | 2004 |
|---|---|---|---|---|
| 11 | 0.000 | 0.000 | 0.000 | 0.000 |
| 12 | 0.000 | 0.000 | 0.000 | 0.000 |
| 13 | 0.003 | 0.002 | 0.000 | 0.000 |
| 14 | 0.013 | 0.008 | 0.004 | 0.020 |
| 15 | 0.038 | 0.026 | 0.040 | 0.036 |
| 16 | 0.083 | 0.079 | 0.102 | 0.105 |
| 17 | 0.154 | 0.149 | 0.179 | 0.143 |
| 18 | 0.191 | 0.220 | 0.257 | 0.249 |
| 19 | 0.254 | 0.249 | 0.268 | 0.307 |
| 20 | 0.304 | 0.291 | 0.299 | 0.302 |
| 21 | 0.278 | 0.287 | 0.313 | 0.334 |
| 22 | 0.282 | 0.224 | 0.306 | 0.303 |
| 23 | 0.284 | 0.227 | 0.303 | 0.328 |
| 24 | 0.269 | 0.237 | 0.297 | 0.296 |
| 25 | 0.279 | 0.231 | 0.260 | 0.283 |
| 26 | 0.237 | 0.171 | 0.264 | 0.301 |
| 27 | 0.237 | 0.223 | 0.318 | 0.298 |
| 28 | 0.231 | 0.211 | 0.213 | 0.263 |
| 29 | 0.204 | 0.221 | 0.241 | 0.266 |
| 30 | 0.222 | 0.180 | 0.268 | 0.260 |
| 31 | 0.181 | 0.168 | 0.192 | 0.225 |
| 32 | 0.206 | 0.234 | 0.237 | 0.239 |
| 33 | 0.190 | 0.202 | 0.233 | 0.189 |
| 34 | 0.152 | 0.242 | 0.251 | 0.191 |
| 35 | 0.224 | 0.137 | 0.225 | 0.213 |
| 36 | 0.191 | 0.152 | 0.148 | 0.222 |
| 37 | 0.129 | 0.117 | 0.218 | 0.183 |
| 38 | 0.138 | 0.136 | 0.145 | 0.191 |
| 39 | 0.143 | 0.095 | 0.168 | 0.115 |
| 40 | 0.119 | 0.128 | 0.071 | 0.135 |
| 41 | 0.114 | 0.068 | 0.106 | 0.064 |
| 42 | 0.074 | 0.045 | 0.069 | 0.067 |
| 43 | 0.063 | 0.061 | 0.100 | 0.072 |
| 44 | 0.030 | 0.074 | 0.042 | 0.075 |
| 45 | 0.036 | 0.035 | 0.024 | 0.030 |
| 46 | 0.055 | 0.066 | 0.104 | 0.039 |
| 47 | 0.000 | 0.011 | 0.029 | 0.019 |
| 48 | | 0.000 | 0.030 | 0.000 |
| 49 | | | | |
| **Total Fertility** | **5.61** | **5.20** | **6.32** | **6.36** |

***Step 5: Derive the age-specific fertility rates***

Single-year age-specific fertility rates for each calendar year are derived by dividing the births in Table 12.2 by the person-years exposed-to-risk in Table 12.3. The results are shown in Table 12.4.

The data vary a lot between calendar years, with estimates of total fertility differing by more than a child per woman between 2002 and 2003. The estimate of total fertility in 2004, despite being derived from only partial exposure in that year for most women, is highly consistent with the estimate for 2003. The shape of the distribution (as can be seen in Figure 12.1) is consistent across the three years, even measured in single years of age. This is true despite a high degree of variability in the estimates by single years of age even if they are aggregated over the three years from 2001 to 2003.

Further aggregating the data into conventional five-year age groups produces the results shown in Table 12.5.

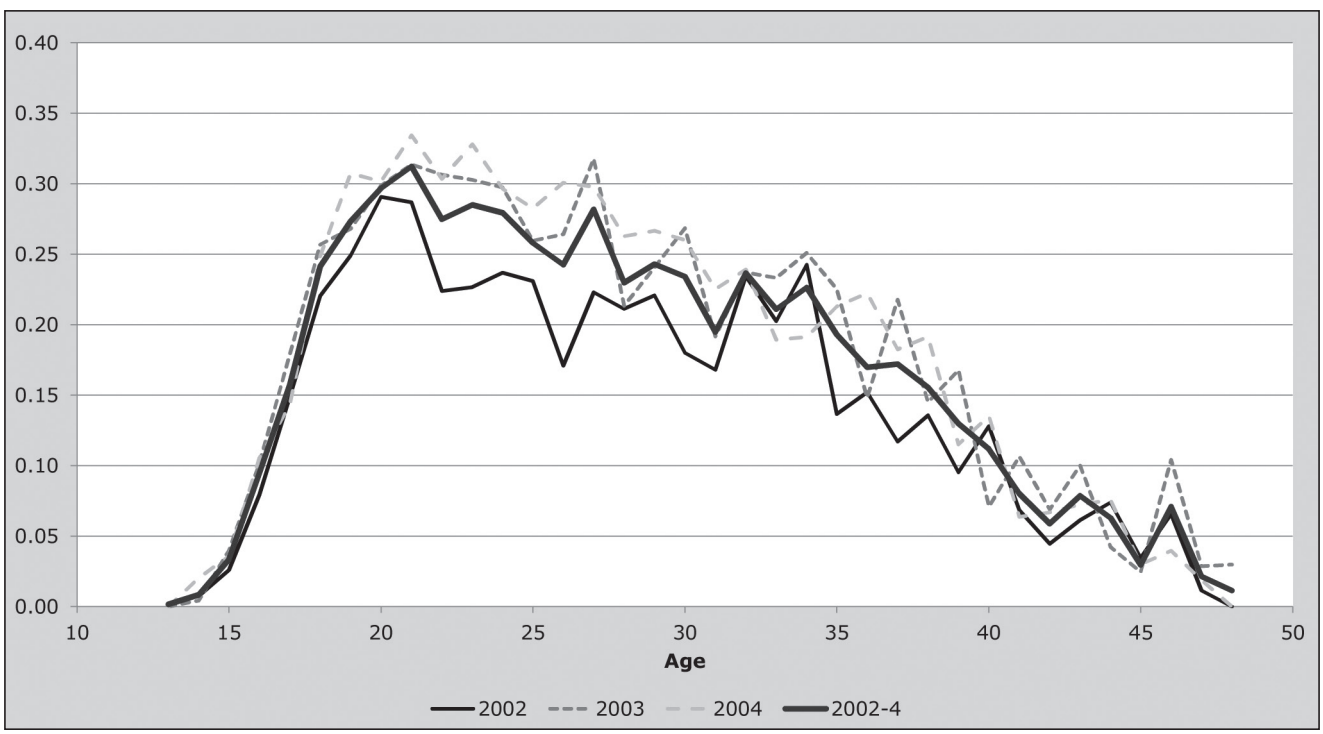
The differences in the last two columns between the ASFRs derived here and those reported in the DHS survey are very small. However, the much lower fertility rates for 2002 (and 2001, not shown) should give cause for concern about possible reference period errors and shifting of births.

*TABLE 12.5 Age-specific fertility rates by grouped year of age and calendar year, Malawi, 2004 DHS*

| Age group | 2002 | 2003 | 2004 | 2002–4 | DHS |
|---|---|---|---|---|---|
| 15–19 | 0.151 | 0.180 | 0.178 | 0.169 | 0.162 |
| 20–24 | 0.254 | 0.304 | 0.312 | 0.290 | 0.293 |
| 25–29 | 0.210 | 0.261 | 0.283 | 0.252 | 0.254 |
| 30–34 | 0.204 | 0.235 | 0.222 | 0.221 | 0.222 |
| 35–39 | 0.129 | 0.180 | 0.184 | 0.164 | 0.163 |
| 40–44 | 0.075 | 0.078 | 0.086 | 0.080 | 0.080 |
| 45–49 | 0.042 | 0.049 | 0.021 | 0.036 | 0.035 |
| **Total Fertility** | **5.32** | **6.44** | **6.43** | **6.05** | **6.05** |

Note: DHS rates are 3-year rates as presented in the 2004 DHS report.
Source: MeasureDHS StatCompiler

*FIGURE 12.1 Age-specific fertility rates by single years of age and calendar year, Malawi, 2004 DHS*



## REFERENCES


Cleland J. 1996. "Demographic data collection in less developed countries", *Population Studies* **50**(3):433–450. doi: 10.1080/0032472031000149556

Rutstein S and G Rojas. 2003. *Guide to DHS Statistics*. Calverton, MD: ORC Macro.

Schoumaker B. 2010. "Reconstructing fertility trends in sub-Saharan Africa by combining multiple surveys affected by data quality problems", Paper presented at Population Association of America 2010 Annual Meeting. Dallas, TX, April 15–17, 2010.

Schoumaker B. 2011. "Omissions of births in DHS birth histories in sub-Saharan Africa: Measurement and determinants", Paper presented at Population Association of America 2011 Annual Meeting. Washington DC, March 31–April 2, 2011.

Schoumaker B. 2013. "A Stata module for computing fertility rates and TFRs from birth histories: tfr2", *Demographic Research* **28**(Article 38):1093–1144. doi: 10.4054/DemRes.2013.28.38

# Chapter 13 The use of *P/F* ratio methods with survey data: Cohort-period fertility rates

*Tom A Moultrie*

## DESCRIPTION OF METHOD

The availability of detailed demographic and birth history data typically collected in demographic surveys (examples being the World Fertility Surveys conducted in the 1970s, and the ongoing programme of Demographic and Health Surveys conducted by ORC Macro) has meant that – in general – direct measures of fertility estimation are to be preferred over indirect methods. Nevertheless, extensions of indirect methods to situations where there is more data can provide not only corroborating evidence to support the results derived directly, but also provide important insights into the quality of the birth history data collected.

One such extension is to apply the same logic as the Brass *P*/*F* ratio method to the birth histories, allowing a detailed investigation of the fertility data by age, period and cohort. The method yields period estimates of total fertility (TF) for either the five-year period or the two five-year periods preceding collection of the data. The method also permits the identification of common errors in the data.

## DATA REQUIREMENTS AND ASSUMPTIONS

### Tabulations of data required

- Numbers of women by age group at the survey date.
- Numbers of births by (current) age group of mother, grouped into five-year periods before the survey date. This tabulation requires, for each entry in a birth history,
  - the date (month and year) of the interview;
  - the child's date of birth (month and year); and
  - the mother's current age group.

### Assumptions

There is no differential fertility between women interviewed in the survey and those who have died or emigrated, and who are therefore not sampled in the survey.

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

The method can be applied working either with the date-handling routines available in most statistical packages, or (almost as well) using dates presented in the DHS CMC format. If dates of birth and interview have not been coded as such, it is recommended that they are so coded for purposes of applying this method. The routine for doing so is described on the MeasureDHS website.

## APPLICATION OF METHOD

The method is applied in the following stages.

### *Step 1: Extract the (weighted) number of women, by age group at survey date*

This is a straightforward tabulation. In the context of DHS data, women's age group at the survey is given by the *v013* variable, and the sample and design weights by *v005* (divided by $10^6$ where appropriate). The number of women in age group $i$ is denoted by $N_i = {}_5N_x$, where $x = 15,20,\ldots, 45$ and $i = x/5–2$.

### *Step 2: Extract numbers of births by age group and period before the survey*

If working with dates in CMC format, with a full birth history in a file with one record per child, the child's current age in months is easily determined by subtracting the CMC of the child's date of birth from the CMC of the interview date. Dividing the result by 60 and taking the integer portion of the result produces an index that allocates the child's date of birth to successive five-year periods before the survey.

A minor modification needs to be made to accommodate the cases where the child was born in the month of interview exactly 5, 10 … years previously. Depending on the relative

timing of the day of interview and the day of birth, children could be in one of two adjacent age groups. To resolve this, and to avoid allocating all such cases to one group, children in the boundary months should be allocated to age groups based on the reported day of interview, if available, and assuming that days of birth are uniformly distributed over each month. Where possible, we define $b = 1$ if day (within the calendar month) of interview < 16 – in other words, the child's birthday is more likely to be in the second half of the month – and 0 otherwise.

Thus

$$Time\ of\ birth = j = \text{int}(\frac{DoI - DoB^c - b}{60}); \quad j = 0,1,2... \tag{1}$$

where *DoI* is the date of the interview and $DoB^c$ is the child's date of birth, both recorded in CMC format. In the case of DHS data, *DoI* is provided by the *v008* variable and $DOB^c$ by variable *b3*. The day of interview is given by variable *v016*.

A cross-tabulation (weighted, where appropriate, for the sample design) of mothers' age group at the survey date and the grouped time of birth variable defined above is then extracted. The structure of the cross-tabulation is as shown in Table 13.1, where the $B_{i,j}$ reflect the aggregate (weighted) number of births $j$ years ago to women in age group $i$ at the survey date:

*TABLE 13.1 Structure of tables used to derive Cohort-Period Fertility Rates*

| | | Births by period before the survey ($j$) | | | | |
|---|---|---|---|---|---|---|
| Age group of cohort at survey ($i$) | Number of women | 0–4 ($j$=0) | 5–9 ($j$=1) | 10–14 ($j$=2) | 15–19 ($j$=3) | 20–24 ($j$=4) |
| 15–19 ($i$=1) | $N_1$ | $B_{1,0}$ | $B_{1,1}$ | | | |
| 20–24 (2) | $N_2$ | $B_{2,0}$ | $B_{2,1}$ | $B_{2,2}$ | | |
| 25–29 (3) | $N_3$ | $B_{3,0}$ | $B_{3,1}$ | $B_{3,2}$ | $B_{3,3}$ | |
| 30–34 (4) | $N_4$ | $B_{4,0}$ | $B_{4,1}$ | $B_{4,2}$ | $B_{4,3}$ | $B_{4,4}$ |
| 35–39 (5) | $N_5$ | $B_{5,0}$ | $B_{5,1}$ | $B_{5,2}$ | $B_{5,3}$ | $B_{5,4}$ |
| 40–44 (6) | $N_6$ | $B_{6,0}$ | $B_{6,1}$ | $B_{6,2}$ | $B_{6,3}$ | $B_{6,4}$ |
| 45–49 (7) | $N_7$ | $B_{7,0}$ | $B_{7,1}$ | $B_{7,2}$ | $B_{7,3}$ | $B_{7,4}$ |

Note that, going back in time, the fertility rates of the youngest women will be zero for time periods in which all the women are less than 10 years old. Some births that occurred in the past will not be reported if the birth histories were collected only from women under 50.

***Step 3: Derive cohort-period fertility rates based on the age group of mother at the time of the survey***

If we denote age groups (or cohorts, defined by age at the survey) by the index, $i$, ($i = 1$ corresponding to the 15–19 age group, etc.) and successive five-year periods before the survey by $j$ ($j = 0$ corresponding to the five-year period immediately preceding the survey, and ending at the survey date), the cohort-period fertility rate is then defined as

$$f_{i,j} = \frac{1}{5}\left(\frac{B_{i,j}}{N_i}\right).$$

The ratio is divided by five because women's exposure will be exactly five years as all women alive at the survey date must have been alive throughout each of the previous periods.

The resulting cohort-period rates present the experience of women in the same cohort (born in the same time period) in the rows, with periods in the columns, and equivalent attained ages running down the diagonals, as shown in Table 13.2.

*TABLE 13.2 Cohort-period fertility rates, classified by age of cohort at survey*

| | Births by period before the survey ($j$) | | | | |
|---|---|---|---|---|---|
| Age group of cohort at survey ($i$) | 0–4 ($j$=0) | 5–9 ($j$=1) | 10–14 ($j$=2) | 15–19 ($j$=3) | 20–24 ($j$=4) |
| 15–19 ($i$=1) | $f_{1,0}$ | $f_{1,1}$ | | | |
| 20–24 (2) | $f_{2,0}$ | $f_{2,1}$ | $f_{2,2}$ | | |
| 25–29 (3) | $f_{3,0}$ | $f_{3,1}$ | $f_{3,2}$ | $f_{3,3}$ | |
| 30–34 (4) | $f_{4,0}$ | $f_{4,1}$ | $f_{4,2}$ | $f_{4,3}$ | $f_{4,4}$ |
| 35–39 (5) | $f_{5,0}$ | $f_{5,1}$ | $f_{5,2}$ | $f_{5,3}$ | $f_{5,4}$ |
| 40–44 (6) | $f_{6,0}$ | $f_{6,1}$ | $f_{6,2}$ | $f_{6,3}$ | $f_{6,4}$ |
| 45–49 (7) | $f_{7,0}$ | $f_{7,1}$ | $f_{7,2}$ | $f_{7,3}$ | $f_{7,4}$ |

***Step 4: Transpose the Cohort-Period Fertility Rates***

The rates derived in Step 3 can also be classified by the age of the mother at the end of each successive five-year time period. The end of the period reflecting births that occurred in the five years before the survey date (when $j = 0$) is the survey date, and the end of the period 5–9 years before the survey (when $j = 1$) is the point exactly five years before the survey. The effect of this reclassification is that a revised series of cohort fertility indices are created:

$$f^*_{k,j} = f_{k+j,j}.$$

With this reclassification, Table 13.2 above is rearranged as in Table 13.3 below. Thus, for example, the fertility of women aged 30–34 at the survey in the period 10–14 years before the survey (i.e. $f_{4,2}$) would now be recast as the fertility of women who were aged 20–24 10 years before the survey ($f^*_{2,2}$).

*TABLE 13.3 Matrix of cohort-period fertility rates, with redefined age*

| | Births by period before the survey ($j$) | | | | |
|---|---|---|---|---|---|
| **Age group of cohort at the end of each period, $k$** | **0–4 ($j$=0)** | **5–9 ($j$=1)** | **10–14 ($j$=2)** | **15–19 ($j$=3)** | **20–24 ($j$=4)** |
| 15–19 ($k$=1) | $f^*_{1,0}$ | $f^*_{1,1}$ | $f^*_{1,2}$ | $f^*_{1,3}$ | $f^*_{1,4}$ |
| 20–24 (2) | $f^*_{2,0}$ | $f^*_{2,1}$ | $f^*_{2,2}$ | $f^*_{2,3}$ | $f^*_{2,4}$ |
| 25–29 (3) | $f^*_{3,0}$ | $f^*_{3,1}$ | $f^*_{3,2}$ | $f^*_{3,3}$ | $f^*_{3,4}$ |
| 30–34 (4) | $f^*_{4,0}$ | $f^*_{4,1}$ | $f^*_{4,2}$ | $f^*_{4,3}$ | |
| 35–39 (5) | $f^*_{5,0}$ | $f^*_{5,1}$ | $f^*_{5,2}$ | | |
| 40–44 (6) | $f^*_{6,0}$ | $f^*_{6,1}$ | | | |
| 45–49 (7) | $f^*_{7,0}$ | | | | |

Periods further back in time will have a steadily rising number of missing values for older women if birth histories were not collected for women over 50. For example $f^*_{6,3}$ would represent the fertility experienced 15–19 years ago by women who were aged 40–44 exactly 15 years before the survey date. At the survey date these women would have been aged 55–59 and would not have completed the birth history section of a DHS survey.

***Step 5: Derive measures of cohort fertility***

Define $P_{k,j}$ to be the cumulated cohort fertility (i.e. attained mean children ever born) from age 15 to the end of age group $k$ of the cohort of women aged $k$ at time $j$, then

$$P_{k,j} = 5.\sum_{z=0}^{k-1} f^*_{k-j,z+j} = 5.\sum_{z=0}^{k-1} f_{k+j,z+j}.$$

***Step 6: Derive measures of period fertility and two estimates of Total Fertility***

Period fertility measures are the cumulated fertility rates in a given period. Thus, we define $F_{i,j}$ to be the cumulated period fertility up to age $i$ in period $j$. Hence,

$$F_{k,j} = 5.\sum_{z=1}^{k} f^*_{z,j} = 5.\sum_{z=1}^{k} f_{z+j,j}.$$

Note that $F_{7,0}$ is a measure of the (period) Total Fertility (TF) in the five years immediately preceding the survey. This estimate can be assumed to apply (roughly) 2½ years before the survey date.

More often than not, $F_{7,1}$ cannot be evaluated directly as this would require reports of fertility among women now aged 50–54 when they were aged 45–49 in the five-year period ending five years before the survey. However, fertility in this age group is generally very low, so an approximate estimate of fertility in the period 5–9 years before the survey can be derived from

$$TF_1 = F_{6,1} + 5.f^*_{7,0}.$$

In other words, we assume that $f^*_{7,1} = f^*_{7,0} = f_{7,0}$. If fertility has been declining, the resulting estimate will be marginally too low, but as fertility is typically very low in this age group, this bias will be unimportant.

In populations with low or moderately low fertility (current total fertility below 3 births per woman), it would be reasonable to make a similar substitution for the unmeasured fertility of women aged 40–49 exactly 10 years before the survey, as fertility in the age group 40–44 would also be low enough for small changes to impact very little on the estimated TF 10–14 years earlier. In this case we could assume that $f^*_{7,2} = f_{7,0}$ and $f^*_{6,2} = f_{6,1}$ to obtain

$$TF_2 = F_{5,2} + 5 \cdot f^*_{7,0} + 5 \cdot f^*_{6,1}.$$

***Step 7: Derive P/F ratios***

The method allows the direct calculation of $P/F$ ratios from the results produced in Steps 5 and 6. The $P/F$ ratio applicable to age group $k$ in period $j$ is

$$\frac{P}{F}(k,j) = \frac{P_{k,j}}{F_{k,j}} = \frac{5 \cdot \sum_{z=0}^{k-1} f^*_{k-z,z+j}}{5 \cdot \sum_{z=1}^{k} f^*_{z,j}} = \frac{5 \cdot \sum_{z=0}^{k-1} f_{k+j,z+j}}{5 \cdot \sum_{z=1}^{k} f_{z+j,j}}.$$

Thus, for example, the $P/F$ ratio for women aged 25–29 in the five year period ending 10 years before the survey is, with $k = 3$ and $j = 2$ in the formulae above,

$$\frac{P}{F}(3,2) = \frac{P_{3,2}}{F_{3,2}} = \frac{\sum_{z=0}^{2} f^*_{3-z,2+z}}{\sum_{z=1}^{3} f^*_{z,2}} = \frac{f^*_{3,2} + f^*_{2,3} + f^*_{1,4}}{f^*_{1,2} + f^*_{2,2} + f^*_{3,2}}$$

$$= \frac{\sum_{z=0}^{2} f_{5,2+z}}{\sum_{z=1}^{3} f_{2+z,2}} = \frac{f_{5,2} + f_{5,3} + f_{5,4}}{f_{3,2} + f_{4,2} + f_{5,2}}.$$

## INTERPRETATION AND DIAGNOSTICS

Several important interpretations arise from these results.

### 1) Estimates of period fertility

Step 6 showed how two estimates of fertility, applicable to points in time approximately 2½ and 7½ years before the survey, can be derived. From this, a short-term trend in fertility may be inferred.

### 2) Interpretation of *P*/*F* ratios and timing of the fertility decline

The *P*/*F* ratios derived in Step 7 can give insights into both the nature and timing of a decline in fertility, as well as problems with the quality of the data. Chapter 6 describes the essential features of the *P*/*F* method.

*P*/*F* ratios of, or very close to, 1 at each age in a given period imply that there has been no change in fertility, as period and cohort measures are roughly equal. Fertility decline is indicated by *P*/*F* ratios that increase consistently with age in any given period, rising from a value close to unity for ages under 25. (25 is used because it is difficult for cohort and period fertility for the youngest cohorts to diverge too much.) Thus, if in one period before the survey, *j*, the *P*/*F* ratios are almost constant by age, but in the next period closer to the survey, *j*–1, the *P*/*F* ratios show a clear trend with respect to age, fertility decline began (approximately) at the date dividing the two periods.

A series of low *P*/*F* ratios in a given period followed or preceded by a series of much higher *P*/*F* ratios is indicative of possible displacement of births into the period where the ratios appear low, and out of the period where the ratios appear to be uncharacteristically high. Similarly, a series of *P*/*F* ratios on a major diagonal (i.e. for a particular cohort) that departs uncharacteristically from the overall trend is indicative of age mis-statement by women, or omissions of births if the trend is observed in the oldest age group.

### 3) Assessment of data quality

Examination of the cohort-period fertility rates (i.e. the $f^*$) derived in Step 4 can contribute to assessment of the quality of the data. For example, reading along the rows from right to left shows how fertility in each age group has changed as the date of the survey approached.

Since, in the absence of severe exogenous factors, the expectation is one of orderly and incremental change, deviations from orderliness may reflect reference period errors or other omissions. Three types of reference period errors are argued to be prevalent in the retrospective birth history data collected in surveys.

The first type of reference period error is that attributed to Brass who argued that older women tend to exaggerate the age of their oldest children thereby placing their birth dates further back in time than they actually occurred. This causes the level of fertility for the earliest periods preceding the survey to be over-estimated and more recent fertility to be under-estimated as births are transferred from relatively recent to more distant time periods, thereby exaggerating the apparent drop in fertility. 'Brass effects' are identified by the earliest cohort-period fertility rates for any age group being distinctly higher than those of slightly younger cohorts at the same attained age. This shifting of births back in time also has the effect of making fertility decline appear less than it is in reality in the more recent periods.

The second type of reference period error is that identified by Potter (1977). Women, Potter argues, have a tendency to bring earlier births closer to the date of the interview, but to report recent events correctly. This results in understatement of the level of fertility for the more distant periods preceding the survey, while recent fertility rates are correct and those in the intervening periods are exaggerated. Potter's model is based on two propositions: first, that the "date of an event is recalled less accurately the longer ago the event took place. The second is that if a birth history is elicited by asking questions about live births in the order in which they occurred, then the date a woman attaches to any event other than the first is influenced by the information she has already given about the previous event" (Potter 1977: 341). 'Potter effects' are more likely to occur when women's birth histories are collected in the order that the births occurred than when asked from youngest to oldest birth.

A third type of error of omission arises from the systematic omission of children born just before the survey, or their displacement into an earlier time period. This is brought

about by the enumerator seeking to avoid asking detailed supplementary questions (for example, an anthropometry questionnaire) of children under a certain age (usually five years). Such errors have been well-documented by Cleland (1996) and Schoumaker (2010, 2011). If such omissions or displacements have occurred, the appearance of fertility decline in the period just before the survey will be exaggerated, and the *P*/*F* ratios in that most recent period will show a much greater degree of fertility decline. Some of this decline may be real, but analysts should be alert to the possible effects of this kind of omission or displacement.

The impact of these three effects on fertility measures over time can be represented graphically, as in Figure 13.1.

The line marked 'true fertility' shows the time trend in fertility (total, or age-specific) in a hypothetical population. 'Brass' effects create the impression of higher fertility in the distant past and a slower fertility decline in the 10 or so years before the survey. 'Potter' effects produce systematic exaggerations of fertility in the 5–15 years before the survey, resulting in a mistaken impression of more rapid recent fertility decline. The 'typical pattern' indicates the nature of common distortions in birth history data. Fertility in the most recent period is usually too low, caused by omission (in censuses) or displacement of recent births to avoid modules on anthropometry etc. (in surveys), while fertility in the far distant past is often exaggerated (by means of 'Brass' effects) and visible in the apparent excess of births to very young women in many birth histories.

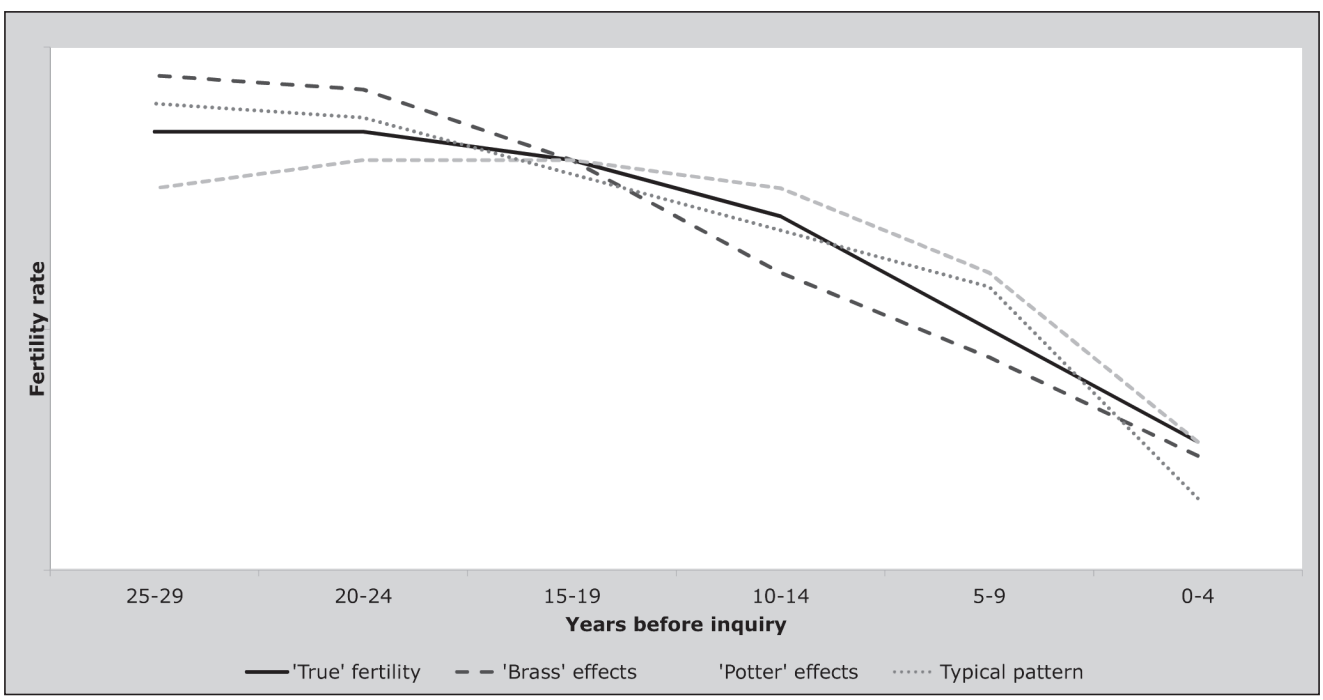
## WORKED EXAMPLE

The example uses retrospective birth history data collected in the 2004 Malawi Demographic and Health Survey. Tabulations have been weighted using the sample and design weights provided with the data, which accounts for the fractional women in the tabulations. The method has been implemented in an accompanying Excel workbook (see website).

### *Steps 1 and 2: Extraction of data*

The tabulations extracted from the DHS data files are shown in Table 13.4. The births reported by period before the survey have been adjusted to accommodate approximately the boundary problem discussed in the derivation of Equation 1.

***FIGURE 13.1*** *Graphical representation of Brass and Potter effects on misreporting of fertility*



***Step 3: Derivation of cohort-period fertility rates, classified by age group of the cohort at the survey date***

Table 13.5 shows the cohort-period fertility rates from the data in Table 13.4. Thus, for example, the cohort fertility rate associated with the births occurring 5–9 years to women aged 20–24 at the survey is given by

$$\frac{1}{5}.\frac{981.8}{2{,}869.7} = 0.068 \ .$$

***Step 4: Transpose the Cohort-Period Fertility Rates***

The rates in Table 13.5 are transposed so that the rows represent equivalent attained ages at the end of each time period represented in the columns. Thus, for example, a woman aged 30–34 at the survey would have been aged 25–29 at the end of the period 5–9 years before the survey (and 20–24 at the end of the period 10–14 years before the survey), and these cohort fertility rates (0.291 and 0.270 from Table 13.5) are now tabulated against ages 25–29 and 20–24 respectively, as in Table 13.6.

To assist with the identification of trends in fertility as well as problems or flaws in the data, the cohort-period fertility rates in Table 13.6 are presented graphically by birth cohort and attained age in Figure 13.2. There would appear to have been some generalized omission of more distant fertility in the survey as the cohort-period fertility rates for the oldest women (the 1955–59 cohort) at younger attained ages (20–24) are somewhat lower than those of slightly younger cohorts. Nonetheless, there are some indications from these data of an incipient fertility decline in Malawi in that fertility rates among the youngest cohorts (those born after 1980) are lower than those of older cohorts.

Further investigations are required to untangle these effects.

***Step 5: Derive measures of cohort fertility***

Cumulated cohort fertility to any given age is calculated by summing the diagonal of cohort rates in Table 13.6 and multiplying by 5, as shown in Table 13.7. Thus, for example, the cumulated cohort fertility of women aged 25–29 at the end of the period 5–9 years before the survey is given by $5(0.291 + 0.270 + 0.081) = 3.210$.

***Step 6: Derive measures of period fertility and two estimates of the TF***

Cumulated period fertility up to a given age are derived by summing all the cohort period fertility rates (CPFRs) from Table 13.6 in a given column (the period) up that age, and multiplying by 5 (shown in Table 13.8). For example, the cumulated fertility up to age 30 in the period 5–9 years before the survey is given by $5(0.068 + 0.250 + 0.291) = 3.047$.

Two estimates of Total Fertility are derived. The first is that in the period 0–4 years before the survey, and is calculated directly from the data (6.1 children per woman). Fertility in the period 5–9 years before the survey is estimated by $6.091 + 5(0.063) = 6.406$ children per woman.

Since the median date of interview in that DHS was November 2004, we can take the two estimates as applying to May 2002 and May 1997 respectively. The results suggest that TF fell by about 0.4 children per woman between the

*TABLE 13.4 Number of women, by age group at survey, and number of births to those women, classified by timing of birth, Malawi, 2004 DHS*

| | | | Births by period before the survey ($j$) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Age group of cohort at survey ($i$) | Approximate cohort birth years | Number of women | 0–4 ($j$=0) | 5–9 ($j$=1) | 10–14 ($j$=2) | 15–19 ($j$=3) | 20–24 ($j$=4) | 25–29 ($j$=5) | 30–34 ($j$=6) |
| 15–19 ($i$=1) | 1985–1989 | 2,392.0 | 713.7 | 5.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 20–24 (2) | 1980–1984 | 2,869.7 | 3,638.8 | 981.8 | 28.7 | 0.0 | 0.0 | 0.0 | 0.0 |
| 25–29 (3) | 1975–1979 | 2,157.4 | 2,952.3 | 2,693.6 | 859.1 | 13.6 | 0.0 | 0.0 | 0.0 |
| 30–34 (4) | 1970–1974 | 1,478.0 | 1,734.4 | 2,152.7 | 1,996.7 | 595.5 | 21.9 | 0.0 | 0.0 |
| 35–39 (5) | 1965–1969 | 1,116.8 | 1,139.6 | 1,462.6 | 1,815.5 | 1,386.4 | 508.4 | 18.1 | 0.0 |
| 40–44 (6) | 1960–1964 | 935.0 | 569.3 | 923.5 | 1,372.6 | 1,456.2 | 1,267.8 | 386.4 | 13.4 |
| 45–49 (7) | 1955–1959 | 749.1 | 235.1 | 558.8 | 952.6 | 1,024.9 | 1,128.3 | 953.3 | 311.6 |

*TABLE 13.5 Cohort period fertility rates, classified by age of mother at the end of each period before the survey, Malawi, 2004 DHS*

| Age group of cohort at survey ($i$) | Period before the survey ($j$) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0–4 ($j$=0) | 5–9 ($j$=1) | 10–14 ($j$=2) | 15–19 ($j$=3) | 20–24 ($j$=4) | 25–29 ($j$=5) | 30–34 ($j$=6) |
| 15–19 ($i$=1) | 0.060 | 0.000 | | | | | |
| 20–24 (2) | 0.254 | 0.068 | 0.002 | | | | |
| 25–29 (3) | 0.274 | 0.250 | 0.080 | 0.001 | | | |
| 30–34 (4) | 0.235 | 0.291 | 0.270 | 0.081 | 0.003 | | |
| 35–39 (5) | 0.204 | 0.262 | 0.325 | 0.248 | 0.091 | 0.003 | |
| 40–44 (6) | 0.122 | 0.198 | 0.294 | 0.311 | 0.271 | 0.083 | 0.003 |
| 45–49 (7) | 0.063 | 0.149 | 0.254 | 0.274 | 0.301 | 0.255 | 0.083 |

*TABLE 13.6 Cohort-period fertility rates, classified by age of mother at the end of each period before the survey, Malawi, 2004 DHS*

| Age group of cohort at the end of each period before the survey, $k$ | Period before the survey ($j$) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0–4 ($j$=0) | 5–9 ($j$=1) | 10–14 ($j$=2) | 15–19 ($j$=3) | 20–24 ($j$=4) | 25–29 ($j$=5) | 30–34 ($j$=6) |
| 15–19 ($k$=1) | 0.060 | 0.068 | 0.080 | 0.081 | 0.091 | 0.083 | 0.083 |
| 20–24 (2) | 0.254 | 0.250 | 0.270 | 0.248 | 0.271 | 0.255 | |
| 25–29 (3) | 0.274 | 0.291 | 0.325 | 0.311 | 0.301 | | |
| 30–34 (4) | 0.235 | 0.262 | 0.294 | 0.274 | | | |
| 35–39 (5) | 0.204 | 0.198 | 0.254 | | | | |
| 40–44 (6) | 0.122 | 0.149 | | | | | |
| 45–49 (7) | 0.063 | | | | | | |

*TABLE 13.7 Cumulative fertility of cohorts at end of each period (P), Malawi, 2004 DHS*

| Age group of cohort at the end of each period before the survey, $k$ | Period before the survey ($j$) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0–4 ($j$=0) | 5–9 ($j$=1) | 10–14 ($j$=2) | 15–19 ($j$=3) | 20–24 ($j$=4) | 25–29 ($j$=5) | 30–34 ($j$=6) |
| 15–19 ($k$=1) | 0.298 | 0.342 | 0.398 | 0.403 | 0.455 | 0.413 | 0.416 |
| 20–24 (2) | 1.610 | 1.647 | 1.754 | 1.697 | 1.769 | 1.689 | |
| 25–29 (3) | 3.015 | 3.210 | 3.322 | 3.327 | 3.195 | | |
| 30–34 (4) | 4.384 | 4.632 | 4.795 | 4.563 | | | |
| 35–39 (5) | 5.652 | 5.783 | 5.835 | | | | |
| 40–44 (6) | 6.391 | 6.581 | | | | | |
| 45–49 (7) | 6.895 | | | | | | |

two periods, although displacement and omissions of most recent births may be producing an exaggerated impression of decline.

***Step 7: Derive P/F ratios***

*P*/*F* ratios are derived by dividing the equivalent cells in Tables 13.7 and 13.8, as shown in Table 13.9.

It is clear from the steady increases in the *P*/*F* ratios with age in the most recent two periods that a fertility decline is under way. No such trend exists in the ratios for 10–14 years before the survey. Thus, the fertility decline in Malawi appears to have started about 10 years before the survey, 1994.

The *P*/*F* ratios offer some evidence that there has been some omission or displacement of births in particular

| Age group of cohort at the end of each period before the survey, *k* | Period before the survey (*j*) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0–4 (*j*=0) | 5–9 (*j*=1) | 10–14 (*j*=2) | 15–19 (*j*=3) | 20–24 (*j*=4) | 25–29 (*j*=5) | 30–34 (*j*=6) |
| 15–19 (*k*=1) | 0.298 | 0.342 | 0.398 | 0.403 | 0.455 | 0.413 | 0.416 |
| 20–24 (2) | 1.566 | 1.591 | 1.749 | 1.644 | 1.811 | 1.686 | |
| 25–29 (3) | 2.935 | 3.045 | 3.375 | 3.202 | 3.317 | | |
| 30–34 (4) | 4.108 | 4.357 | 4.843 | 4.570 | | | |
| 35–39 (5) | 5.129 | 5.345 | 6.115 | | | | |
| 40–44 (6) | 5.738 | 6.091 | | | | | |
| 45–49 (7) | 6.051 | 6.406 | | | | | |

*TABLE 13.8 Cumulative fertility within periods (F), Malawi, 2004 DHS*

cohorts and periods. For example, the data on births occurring 10–14 years before the survey indicates *P*/*F* ratios less than one. In addition, the ratio in respect of births 10–14 years before of women then aged 20–24 (1.003) is clearly divergent from the ratio of women of the same age 5–9 years before the survey and 15–19 years before the survey. This might be attributable to rising fertility in this period, but this seems unlikely. More probably – since the ratios are too low in that period – births have been shifted into that period (possible Brass or Potter effects), thereby inflating the estimates of *F* and depressing the ratios. Women aged 45–49 at the survey would appear to have omitted some of their births in that in all periods 5–25 years before the survey, the *P*/*F* ratios for this cohort are lower than those of the cohort of women aged 40–44 at the survey.

*FIGURE 13.2 Cohort-period fertility rates, Malawi, 2004 DHS*

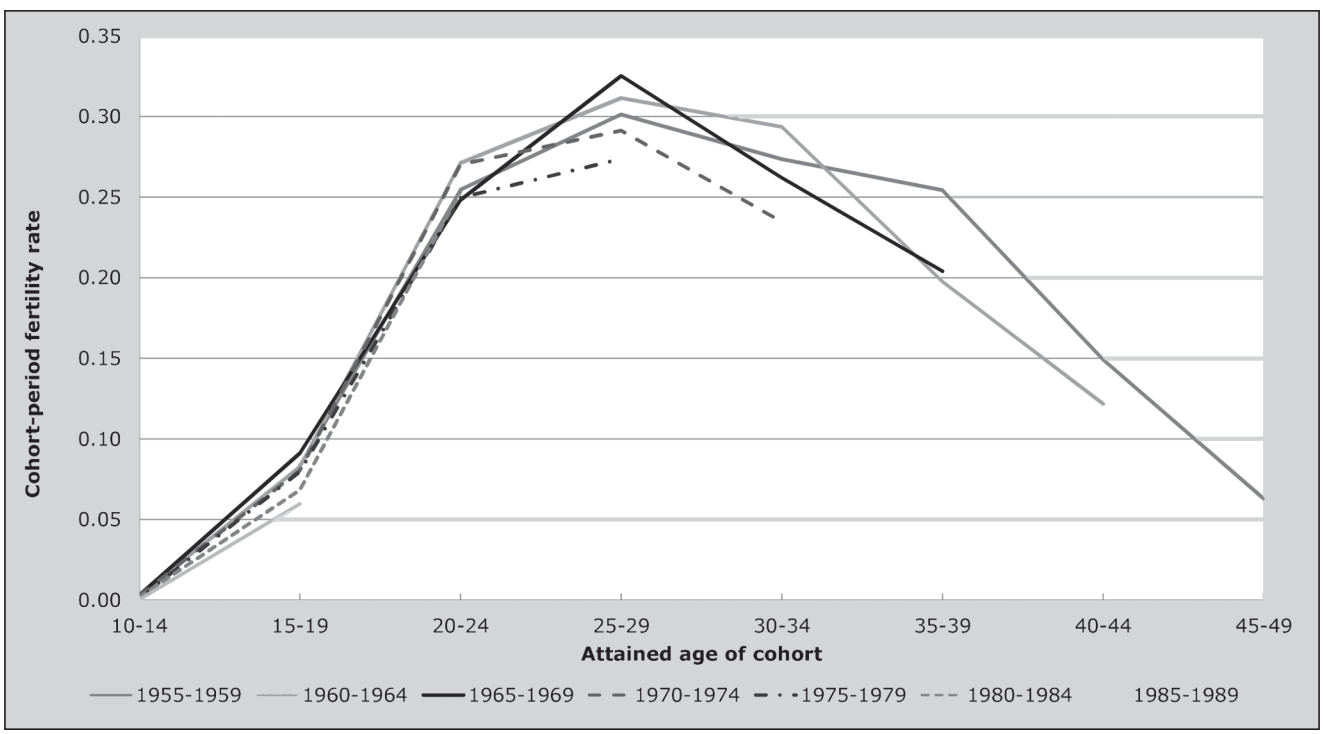


*TABLE 13.9 P/F ratios, Malawi, 2004 DHS*

| Age group of cohort at the end of each period before the survey, *k* | (*j*) 0–4 (*j*=0) | 5–9 (*j*=1) | 10–14 (*j*=2) | 15–19 (*j*=3) | 20–24 (*j*=4) | 25–29 (*j*=5) | 30–34 (*j*=6) |
|---|---|---|---|---|---|---|---|
| 15–19 (*k*=1) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 20–24 (2) | 1.028 | 1.035 | 1.003 | 1.032 | 0.977 | 1.002 | |
| 25–29 (3) | 1.027 | 1.054 | 0.984 | 1.039 | 0.963 | | |
| 30–34 (4) | 1.067 | 1.063 | 0.990 | 0.998 | | | |
| 35–39 (5) | 1.102 | 1.082 | 0.954 | | | | |
| 40–44 (6) | 1.114 | 1.080 | | | | | |
| 45–49 (7) | 1.139 | | | | | | |

## FURTHER READING AND REFERENCES

The method of deriving *P*/*F* ratios from survey data described here was set out in the early 1980s by Hobcraft and others (Goldman and Hobcraft 1982; Hobcraft, Goldman and Chidambaram 1982). Hobcraft, Goldman and Chidambaram (1982), in their exposition of the method, set out the approach to analysing cohort-period fertility rates by duration since marriage (and age at marriage) and duration since first birth (and age at first birth).

Beyond the sources already cited, the method has been used in a number of analyses of World Fertility Survey and DHS data. Examples include applications to Lesotho (Timæus and Balasubramanian 1984), Zimbabwe (Muhwava and Timæus 1996), West Africa (Onuoha and Timæus 1995) and Nepal (Collumbien, Timæus and Acharya 2001). Hinde and Mturi (2000) applied the method, using duration since marriage, to Tanzanian data.


Cleland J. 1996. "Demographic data collection in less developed countries", *Population Studies* **50**(3):433–450. doi: 10.1080/0032472031000149556

Collumbien M, IM Timæus and L Acharya. 2001. "Fertility decline in Nepal," in Sathar, ZA and JF Philips (eds). *Fertility Transition in South Asia*. Oxford: Oxford University Press, pp. 99–120.

Goldman N and J Hobcraft. 1982. *Birth Histories*. WFS Comparative Studies 17. Voorburg, Netherlands: International Statistical Institute.

Hinde A and AJ Mturi. 2000. "Recent trends in Tanzanian fertility", *Population Studies* **54**(2):177–191. doi: 10.1080/713779080

Hobcraft JN, N Goldman and VC Chidambaram. 1982. "Advances in the P/F ratio method for the analysis of birth histories", *Population Studies* **36**(2):291–316. doi: 10.2307/2174202

Muhwava W and IM Timæus. 1996. *Fertility Decline in Zimbabwe*. Centre for Population Studies Research Paper 96–1. London: London School of Hygiene & Tropical Medicine.

Onuoha NC and IM Timæus. 1995. "Has a fertility transition begun in West Africa?", *Journal of International Development* **7**(1):93–116. doi: 10.1002/jid.3380070107

Potter JE. 1977. "Problems in using birth-history analysis to estimate trends in fertility", *Population Studies* **31**(2):335–364. doi: 10.2307/2173921

Schoumaker B. 2010. "Reconstructing fertility trends in sub-Saharan Africa by combining multiple surveys affected by data quality problems", Paper presented at Population Association of America 2010 Annual Meeting. Dallas, TX, April 15–17, 2010.

Schoumaker B. 2011. "Omissions of births in DHS birth histories in sub-Saharan Africa: Measurement and determinants", Paper presented at Population Association of America 2011 Annual Meeting. Washington DC, March 31–April 2, 2011.

Timæus IM and K Balasubramanian. 1984. *Evaluation of the Lesotho Fertility Survey 1977*. WFS Scientific Reports, 58. Voorburg, Netherlands: International Statistical Institute.

# Vital Registration Data Methods

# Chapter 14 Comparison of mean number of births registered by a cohort of women with the reported average parity of the same cohort

*Tom A Moultrie and Basia Zaba*

## DESCRIPTION OF METHOD

The synthetic relational Gompertz model presents a method for comparing average parities with average parity equivalents derived from period fertility rates without having to assume constant fertility. The most important aspect of that method is that average parities are calculated for a period rather than for a series of cohorts. It requires, however, that data on children ever born be available for two points in time, five or ten years apart. If only one source of data on children ever born exists, or if the inter-survey period is not (at least approximately) five or ten years, an alternative procedure that does not require the assumption of constant fertility must be used.

The method outlined here is such a procedure. However, while it has distinct theoretical benefits, in practical application the method requires that a fairly long series of annual data on registered births classified by age of mother is available. Such data may not be readily accessible or may be deemed to be so unreliable and defective that the method may not produce sensible results. The method is mainly of use with data on births from a vital registration system, which is normally the only source of information about births by calendar year over a 20-year period. However, if fertility schedules are available from other sources for regular five-year intervals (e.g. complete birth history data collected at a series of time points), such schedules could be used. The method is described here in terms of data from a vital registration system. In general, if parities up to age $x$ are to be used, current fertility data on at least the previous $x-15$ years are required in order to make reasonable comparisons.

The method makes use of the cohort nature of reported average parities and compares them with parity equivalents obtained from the recorded fertility rates pertaining to the relevant cohorts. If one considers women aged 30–34 at some census, then ten years before the census they were aged 20–24, and 20 years before the census they were aged 10–14. Therefore, on the assumption that childbearing begins at age 15, the children ever born reported by women aged up to 35 at the time of the census reflect the cumulated fertility experience of the women over the preceding 20 years. If mortality and migration are assumed to be unrelated to the fertility experience of women, and fertility rates can be calculated for those 20 years, average parity equivalents for each cohort can be constructed and compared with the reported average parity of women at the time of the census.

The difficulty with applying this general idea is that a cohort represented by a conventional five-year age group at the time of the census would not have been in a conventional five-year age group in each of the earlier years. Thus, the population in age group 30–34 at the time of a census would have been aged 29–33 a year earlier, 28–32 two years earlier and so on. If registered births are tabulated by single year of age of mother, this problem is not serious, because single-year fertility rates can be calculated for each year and then relatively easily summed by cohort. The cross-tabulations and calculations would be lengthy, however, and age-heaping might have a non-trivial effect on the fertility rates. It is therefore convenient to have an approach that can be applied to rates for conventional five-year age groups.

Importantly, the method also finds use as a means of evaluating the completeness of birth registration in a vital registration system. The method described here circumvents many of the interpolation problems described in its initial formulation in *Manual X* (UN Population Division 1983: 45–55) by reformulating the approach as another variant of the relational Gompertz model.

## DATA REQUIRED

The data required for this method are:

- The number of children ever born, or average parities, by five-year age group of mother, taken from a recent census.
- Registered births by five-year age group of mother for each of 15 or 20 years preceding the census.
- The number of women in each age group from the census, and from one or more earlier censuses, to allow the estimation of the female population by five-year age group for each of the 15 or 20 years preceding the final census.

## ASSUMPTIONS

Most of the assumptions are those associated with the relational Gompertz model, namely:

- The standard fertility schedule chosen for use in the fitting procedure appropriately reflects the shape of the fertility distribution in the population.
- Any inter-survey changes in fertility have been smooth and gradual and have affected all age groups in a broadly similar way.
- Errors in the pre-adjustment fertility rates are proportionately the same for women in the central age groups (20–39), so that the age pattern of fertility described by reported births in the past year is reasonably accurate.

The method also makes the assumption that the populations used as denominators in the derivation of fertility rates have been corrected for under-enumeration, or other errors. This is particularly important because the numerator and denominator of the rates come from different sources.

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Before commencing analysis of fertility levels using this method, analysts should investigate the quality of the data at least in the following dimensions:

- age and sex structure of the population;
- reported births in the last year; and
- average parities and whether an el-Badry correction is necessary.

## CAVEATS AND WARNINGS

The objective of this method is to estimate the recent levels of fertility and to measure the completeness of birth registration, with a view to adjusting births registered during a recent period to compensate for omission. The effects of other errors, such as changes in the completeness of census enumeration through time, should therefore be allowed for before cumulating age-specific fertility rates for comparison with average parities. Hence, when there is evidence suggesting that there have been changes in the completeness of enumeration, the censuses should, if at all possible, be adjusted before calculating the population denominators. It is not necessary to adjust each census for absolute under-enumeration; it is only necessary to ensure that the relative completeness of enumeration of the earlier censuses is the same as that of the most recent one.

## APPLICATION OF METHOD

The following steps are required for the computational procedure.

### *Step 1: Calculate the reported average parities*

Calculate the average parities at the last census, ${}_5P_x$, of women in each age group $[x, x+5)$, for $x = 15,20\ldots 45$ if this has not been done already as part of earlier analyses.

### *Step 2: Estimate the mid-year female population by age group for each year preceding the last census*

The exact procedure to be followed in estimating the series of mid-year female populations by age group depends upon the dates of the census enumerations available. The procedure is therefore discussed here in general terms.

It is assumed that census enumerations cover, or almost cover, the years for which registered fertility rates are to be cumulated, since it is preferable that the procedure described below is used only for interpolation of the population between census dates, rather than for extrapolation to dates before or after those covered by the censuses.

Define the reference date, $t_a$, of each census, $a$, taking into account the fraction of the year up to and including the exact date, or central reference date for the census. Thus, for example, the number representing the date of a census conducted on the night of 9–10 October 2001 would be calculated by summing up the days of the year preceding the census night as[1]

$$2001 + \frac{(31+28+31+30+31+30+31+31+30+9)}{365}$$

$$= 2001 + \frac{282}{365} = 2001.773.$$

1 The formula for decimalizing dates in Excel is: YEAR(<cellref>) +YEARFRAC(DATE(YEAR(<cellref>),1,0), <cellref>–1,1) where <cellref> is the cell where the date is captured.

The exponential growth rate, $r(i,a)$, of each age group $i$ between census $a$ and census $a+1$ is then obtained by dividing the difference between the natural logarithms of the female population of age group $i$ at the second and first censuses, $N(i,t_{a+1})$ and $N(i,t_a)$ respectively, by the length of the intercensal period in years:

$$r(i,a) = \frac{\ln\left(\frac{N(i,t_{a+1})}{N(i,t_a)}\right)}{t_{a+1} - t_a} \quad . \qquad (1)$$

If there are three censuses covering the period of investigation, one would derive $r(i,1)$ for the period between the first and second census, and $r(i,2)$ covering the period between the second and third census.

The estimated mid-year population for each year for each age group can then be calculated for each year between $t_a$ and $t_{a+1}$ by expanding exponentially the initial population:

$$N(i,t) = N(i,t_a)\exp\left(r(i,a)(t+0.5-t_a)\right) \quad \text{for } t_a<t<t_{a+1}. \qquad (2)$$

If it is necessary to extrapolate to dates before the first census or after the last one, the growth rate in the closest intercensal interval should be used.

As the $N(i,t)$ are estimated mid-year populations in age group $i$ at time $t$, they represent approximate denominators for the calculation of age-specific fertility rates in that year.

***Step 3: Calculate age-specific fertility rates from births registered during the years preceding the census***

Age-specific fertility rates are required for a total of 15 calendar years less than the upper age for which parities will be used. For example, if parities up to age 35 (i.e. the 30–34 age group) will be used, a series of fertility rates stretching back 20 years is required. The rate for age group $i$ and calendar year $t$, $f(i,t)$, is calculated as

$$f(i,t) = \frac{B(i,t)}{N(i,t)}$$

where $B(i,t)$ is the number of births registered in calendar year $t$ as having occurred to women of age group $I$ in calendar year $t$.

If registered births by age of mother are not available for a few of the 20 calendar years required, the application of the method will be only slightly affected if rates for the missing year(s) are interpolated from neighbouring rates. For example, if fertility rates are only available for the last 16 of the 20 years required to analyse registration completeness for women under age 35, the rates for the earliest available calendar year can be adopted for the four preceding years without much danger of introducing sizeable errors, since the imputed values used in the analysis would pertain to women aged 15–19 at the time, an age range where fertility is generally relatively low. However, it would be unwise to extrapolate data from the past to impute values for the most recent years because the imputation would affect more age groups, including the years of peak childbearing in which fertility changes over time are more likely to take place.

***Step 4: Cumulate registered fertility for different female birth cohorts to estimate parity equivalents***

Birth data are tabulated by calendar year, so age-specific fertility rates will be calculated for calendar years. Cumulating rates to the end of each age group will therefore produce fertility cumulants that correspond to the ends of calendar years. The census providing average parities is unlikely to have as its reference date exactly the end of a year, but since average parities for a specified age group change only slowly even when fertility is changing rapidly, the parities from the census can be regarded as referring to the year-end nearest to the census date, and registered rates can be cumulated up to the nearest year-end. Thus, if the census date is on or before 30 June, registered fertility rates would be cumulated to the end of the preceding calendar year, whereas if the census date is after 30 June, registered fertility rates would be cumulated to the end of the calendar year during which the census took place. More accurate and refined approaches (for example using exponential extrapolation) are unwarranted given the overall uncertainties and errors in the method.

To estimate the current fertility parity equivalents, a relational Gompertz model is applied to the age-specific fertility rates derived in the previous step. First, the age-specific fertility rates in each year are cumulated to ages 20, 25, 30 and 35. Then, as with the standard relational Gompertz model, the gompits of the ratios of adjacent cumulated fertility estimates are taken. These correspond to the $z(x)$ in the conventional formulation of the model.

Second, estimates of the parameters are derived from the values of $z(x)$ calculated for each year. This is done in a manner analogous to the fitting of a relational Gompertz model only to the $F$-points in the conventional formulation of the model. We again assume a 20-year run of registration

data on ages 15–34. A regression of $z(x)$–$e(x)$ against $g(x)$, where $e(x)$ and $g(x)$ are derived from the standard fertility schedule, across the three ages 20, 25 and 30, then allows estimates of $\alpha(t)$ and $\beta(t)$ to be derived. $\beta(t)$ is the slope of the linear regression based on the three values of $z(x)$ in year $t$; the intercept is given by $\alpha(t) + (\beta(t)-1)^2.c/2$, where $c$ is a constant derived from the fertility standard used.

Third, the relational Gompertz model is used to produce estimates of fertility by single ages for each calendar year. The values of $\alpha(t)$ and $\beta(t)$ derived in the previous step are used to determine the shape of the fertility schedule, while the level is derived from the fertility cumulated to each of ages 25, 30 and 35. The estimate of total fertility associated with the fertility cumulated to age $x$ in a given year $t$ is given by

$$TF(x,t) = \frac{F(x,t)}{\exp\left(-\exp\left(-\alpha(t) - \beta(t)Y^s(x)\right)\right)} \qquad (3)$$

where $F(x,t)$ is cumulated fertility to age $x$ in year $t$ and $Y^s(x)$ is the gompit at age $x$ of the modified Zaba standard fertility schedule. The resulting three estimates of total fertility are averaged to provide an estimate of the total fertility in each year. Having solved for $TF(x,t)$, it is a simple matter to estimate the fertility between ages $x$ and $x+1$ at time $t$, $f(x,t)$:

$$f(x,t) = TF(x,t) \cdot \left( \begin{array}{c} \exp\left(-\exp\left(-\alpha(t) - \beta(t)Ys(x+1)\right)\right) \\ -\exp\left(-\exp\left(-\alpha(t) - \beta(t)Ys(x)\right)\right) \end{array} \right). \qquad (4)$$

Fourth, the parity equivalents are derived by summing the single-age, single-year fertility rates for each five-year cohort, and dividing by five (the width of the age interval). If we denote these parity equivalents as $E(i,s)$, where $i=1$ corresponds to the 15–19 age group, $i=2$ to the 20–24 age group etc. and $s$ denotes the last year for which estimates will be derived, then

$$E(i) = \sum_{j=0}^{5i+3} \sum_{m=5i+9}^{5i+13} f(m-j, s-j). \qquad (5)$$

***Step 5: Estimate the completeness of birth registration***

The cumulated cohort fertility from registered births, $E(i)$, calculated in the previous step has been constructed so as to be comparable to reported cohort parity, $P(i)$, at the final census. Therefore, the ratio $E(i)/P(i)$ provides a measure of the average completeness of registration of the births that occurred to cohort $i$. If the completeness of registration had remained approximately constant over a period of 15 years or so, the $E/P$ ratios should have more or less the same values for all cohorts, and an average of the ratios for age groups 20–24, 25–29 and 30–34 can be used as an estimate of the completeness of birth registration over the period. Its reciprocal can be used as an adjustment factor for any or all of the age-specific fertility schedules calculated in Step 3.

Two forces are in opposition in the interpretation of these $E/P$ ratios. First, in general, if the completeness of birth registration has been improving over time, the $E/P$ ratios for the younger cohorts will be higher than for older cohorts. In such a situation, the most recent fertility schedule (based on the registered births) may be adjusted by $P(2)/E(2)$, the ratio reflecting the most recent level of completeness. ($P(1)/E(1)$ should usually not be used as an adjustment factor because of the intrinsic difficulty in approximating $E(1)$ accurately.) However, when the $E/P$ ratios indicate that completeness has been improving over time, no obvious basis exists for adjusting the fertility schedules referring to earlier years.

The second force arises because the $E/P$ ratios may be lower for younger women than for older women as a result of less reliable and punctual registration of births among younger mothers. If this is the case, choosing the $E/P$ ratio at a younger age will give an inaccurate portrayal of the completeness of the vital registration data. By contrast, the estimates of completeness based on the reports of women aged 25–29 and 30–34 are affected relatively little by excess omission at early ages, so a case could be made for using the average of these two $E/P$ ratios to determine completeness.

## INTERPRETATION AND DIAGNOSTICS

An important assumption of this method is that the denominators used to calculate the age-specific fertility rates are accurate. Because the numerator and denominator for the estimated age-specific fertility rates do not come from the same source, age-reporting errors that affect birth registration and population enumeration differently will distort the pattern of period age-specific fertility rates. Erratic variation in the age-specific growth rates might suggest problems with age-reporting, but the effects of age misreporting on the final estimates of completeness are very hard to predict. Denominators may also be distorted by changes in the completeness of enumeration from one census to the next and differential completeness of enumeration by age group might affect the

results. Changes in enumeration completeness might also affect average parities. For example, if women with children are more likely to be enumerated than women without, average parities will be inflated by omission.

## WORKED EXAMPLE

The example presented here uses data from three censuses conducted in Chile in April 1970, April 1982 and April 1992 respectively. The original tabulations were taken from the United Nations Statistics Division Demographic Yearbooks, using tabulations on number of women by age group at various censuses; reported births by age of mother and year; mean children ever born at the last census. The method has been implemented in an accompanying Excel workbook (see website).

### *Step 1: Calculate the reported average parities for ages 15–19 … 30–34 from the final census*

The average parities by age from the 1992 Census shown in Table 14.1 are those presented in the Demographic Yearbook.

*TABLE 14.1 Average parities, Chile, 1992 Census*

| Age group | 15–19 | 20–24 | 25–29 | 30–34 |
|---|---|---|---|---|
| Average parity | 0.14 | 0.69 | 1.37 | 2.02 |

No check is possible on these data as the underlying tabulations are not readily available.

### *Step 2: Estimate the mid-year female population by age group for each year preceding the last census*

Table 14.2 gives the numbers of women by age group enumerated in each of the three censuses, as downloaded from the Demographic Yearbook.

The reference dates for the censuses are 1970.304, 1982.304 and 1992.306. (The reference date for the 1992 census reference date is slightly different from the other two because 1992 is a leap year. The reference date is calculated as $1992 + (31 + 29 + 31 + 21)/366 = 1992.306$, on the assumption that the census date refers to the night of the 21–22 April in each case.

The growth rate in each intercensal period is then derived using Equation 1 as shown in Table 14.3.

*TABLE 14.2 Numbers of women by age group enumerated in the 1970, 1982 and 1992 Chile Censuses*

| | | Age group | | | |
|---|---|---|---|---|---|
| *a* | Census date ($t_a$) | 15–19 | 20–24 | 25–29 | 30–34 |
| 1 | 22-Apr-70 | 466,736 | 398,383 | 324,130 | 267,312 |
| 2 | 22-Apr-82 | 652,552 | 595,598 | 479,199 | 399,344 |
| 3 | 22-Apr-92 | 600,563 | 608,933 | 623,305 | 576,710 |

*TABLE 14.3 Intercensal age-specific growth rates, Chile, 1970, 1982 and 1992 Censuses*

| | Age group | | | |
|---|---|---|---|---|
| Intercensal period | 15–19 | 20–24 | 25–29 | 30–34 |
| 1970–1982 | 0.0279 | 0.0335 | 0.0326 | 0.0335 |
| 1982–1992 | –0.0083 | 0.0022 | 0.0263 | 0.0367 |

Thus, for example, the growth rate in the 30–34 age group between the 1982 and 1992 Censuses is given by

$$r(3,2) = \frac{\ln\left(\frac{576,710}{399,344}\right)}{1992.306 - 1982.304} = 0.0367.$$

The growth rates tend to increase with age group and decline over time. This is suggestive of a declining fertility pattern that may have begun some decades earlier, reinforcing the need for an analytical method that does not assume unchanging fertility.

On the basis of these growth rates, the estimated mid-year population of women by age group can be derived for each calendar year using Equation 2. The resulting values of $N(i,t)$ are shown in Table 14.4.

For example the population of 15–19 year old women in mid-1990 is given by

$$N(1,1990) = 652,552 \exp\left(-0.0083(1990.5 - 1982.3)\right)$$
$$= 609,634.$$

### *Step 3: Calculate age-specific fertility rates from births registered during the years preceding the census*

The number of births reported by age group and year, and downloaded from the Demographic Yearbook is shown in Table 14.5.

Age-specific fertility rates are derived by dividing the births (Table 14.5) by the estimated mid-year population for each age group and year (Table 14.4). The results are shown in Table 14.6.

*TABLE 14.4 Estimated mid-year populations by age group, Chile*

| | Age group | | | |
|---|---|---|---|---|
| Year | 15–19 | 20–24 | 25–29 | 30–34 |
| 1972 | 496,255 | 428,806 | 348,169 | 287,686 |
| 1973 | 510,309 | 443,420 | 359,700 | 297,472 |
| 1974 | 524,761 | 458,532 | 371,612 | 307,591 |
| 1975 | 539,623 | 474,159 | 383,919 | 318,054 |
| 1976 | 554,906 | 490,318 | 396,633 | 328,873 |
| 1977 | 570,621 | 507,029 | 409,769 | 340,060 |
| 1978 | 586,781 | 524,308 | 423,340 | 351,628 |
| 1979 | 603,400 | 542,177 | 437,359 | 363,589 |
| 1980 | 620,488 | 560,655 | 451,844 | 375,957 |
| 1981 | 638,061 | 579,762 | 466,808 | 388,746 |
| 1982 | 651,492 | 599,521 | 482,267 | 401,969 |
| 1983 | 646,106 | 597,177 | 494,503 | 417,283 |
| 1984 | 640,765 | 598,500 | 507,674 | 432,902 |
| 1985 | 635,469 | 599,827 | 521,196 | 449,104 |
| 1986 | 630,215 | 601,156 | 535,078 | 465,913 |
| 1987 | 625,006 | 602,489 | 549,331 | 483,352 |
| 1988 | 619,839 | 603,824 | 563,962 | 501,443 |
| 1989 | 614,715 | 605,162 | 578,984 | 520,211 |
| 1990 | 609,634 | 606,503 | 594,405 | 539,681 |
| 1991 | 604,595 | 607,847 | 610,238 | 559,880 |

*TABLE 14.5 Reported births by age group of mother and year, Chile*

| | Age group | | | |
|---|---|---|---|---|
| Year | 15–19 | 20–24 | 25–29 | 30–34 |
| 1972 | 39,839 | 80,430 | 64,624 | 38,937 |
| 1973 | 40,241 | 82,108 | 63,949 | 38,499 |
| 1974 | 39,884 | 79,316 | 63,477 | 37,880 |
| 1975 | 39,086 | 75,519 | 59,365 | 35,863 |
| 1976 | 37,658 | 73,889 | 57,171 | 34,129 |
| 1977 | 36,104 | 71,445 | 53,467 | 32,190 |
| 1978 | 37,138 | 73,224 | 53,725 | 31,832 |
| 1979 | 36,833 | 75,905 | 55,361 | 32,537 |
| 1980 | 38,562 | 79,724 | 59,771 | 33,769 |
| 1981 | 40,252 | 86,037 | 64,849 | 36,494 |
| 1982 | 39,298 | 86,061 | 68,029 | 38,406 |
| 1983 | 36,077 | 81,213 | 65,236 | 37,506 |
| 1984 | 37,571 | 83,960 | 67,266 | 39,105 |
| 1985 | 34,946 | 80,735 | 69,180 | 39,828 |
| 1986 | 35,925 | 83,434 | 72,876 | 42,605 |
| 1987 | 35,633 | 84,674 | 75,416 | 45,037 |
| 1988 | 37,354 | 87,484 | 80,527 | 48,290 |
| 1989 | 39,095 | 86,990 | 82,919 | 50,875 |
| 1990 | 39,543 | 85,292 | 84,336 | 52,942 |
| 1991 | 38,324 | 79,406 | 81,907 | 53,425 |

***Step 4: Cumulate registered fertility for different female birth cohorts to estimate parity equivalents***

The derivation of the parity equivalents proceeds as follows. The age-specific fertility rates in Table 14.6 are cumulated to the upper limit of each age group. Thus, for example, the cumulated fertility to age 25 in 1972 would be calculated from $5(0.0803 + 0.1876) = 1.3392$. The cumulated fertility to age 30 in 1972 is $5(0.0803 + 0.1876 + 0.1856) = 2.2673$. The cumulated rates are shown in Table 14.7.

For each year, the ratio of cumulated fertility in a given age group to that in the next oldest age group is calculated. Thus in the example in the previous paragraph, the ratio of cumulated fertility at age 25 to that at age 30 would be $1.3392/2.2673 = 0.5907$. The ratios are shown in the first three columns of Table 14.8.

Next, a gompit (double negative-log) transform is applied to the ratios to produce a value of $z(x)$ for each of ages 25, 30 and 35 in each year. Using the same example, the value of $z(25)$ in 1972 would be $-\ln(\ln(0.5907)) = 0.6415$. The gompits are shown in the last three columns of Table 14.8.

Values of $e(x)$ and $g(x)$ are tabulated without an age-shift since the data on fertility comes from a vital registration system and hence reflects the age of the mother at the birth of the child. The values are derived from Zaba's modified version of the Booth fertility standard, the only peer-reviewed standard for women currently available. (The standard, and the process of deriving $e(x)$ and $g(x)$, are described in detail in the description of the relational Gompertz method.) The values of $e(x)$ and $g(x)$ for the ages required to fit a relational Gompertz model to the observed fertility data are shown in Table 14.9.

In each year, therefore, it is possible to derive values of $z(x)–e(x)$ for three ages, and to regress these values against the tabulated values of $g(x)$ for the same ages. The value of $\beta$ in the relational Gompertz model is the slope of the regression equation, while $\alpha$ is estimated from

$$\alpha = \text{intercept} - (\beta - 1)^2 . \frac{c}{2}$$

where $c$ is a constant (=0.95739) derived from the modified Zaba fertility standard. The resulting values of $\alpha$ and $\beta$ for each year are shown in the first two columns of Table 14.10.

*TABLE 14.6 Age-specific fertility rates by age group and year, Chile*

| | Age group | | | |
|---|---|---|---|---|
| Year | 15–19 | 20–24 | 25–29 | 30–34 |
| 1972 | 0.0803 | 0.1876 | 0.1856 | 0.1353 |
| 1973 | 0.0789 | 0.1852 | 0.1778 | 0.1294 |
| 1974 | 0.0760 | 0.1730 | 0.1708 | 0.1232 |
| 1975 | 0.0724 | 0.1593 | 0.1546 | 0.1128 |
| 1976 | 0.0679 | 0.1507 | 0.1441 | 0.1038 |
| 1977 | 0.0633 | 0.1409 | 0.1305 | 0.0947 |
| 1978 | 0.0633 | 0.1397 | 0.1269 | 0.0905 |
| 1979 | 0.0610 | 0.1400 | 0.1266 | 0.0895 |
| 1980 | 0.0621 | 0.1422 | 0.1323 | 0.0898 |
| 1981 | 0.0631 | 0.1484 | 0.1389 | 0.0939 |
| 1982 | 0.0603 | 0.1435 | 0.1411 | 0.0955 |
| 1983 | 0.0558 | 0.1360 | 0.1319 | 0.0899 |
| 1984 | 0.0586 | 0.1403 | 0.1325 | 0.0903 |
| 1985 | 0.0550 | 0.1346 | 0.1327 | 0.0887 |
| 1986 | 0.0570 | 0.1388 | 0.1362 | 0.0914 |
| 1987 | 0.0570 | 0.1405 | 0.1373 | 0.0932 |
| 1988 | 0.0603 | 0.1449 | 0.1428 | 0.0963 |
| 1989 | 0.0636 | 0.1437 | 0.1432 | 0.0978 |
| 1990 | 0.0649 | 0.1406 | 0.1419 | 0.0981 |
| 1991 | 0.0634 | 0.1306 | 0.1342 | 0.0954 |

*TABLE 14.7 Cumulated fertility to age x, by year, Chile*

| | Age group | | | |
|---|---|---|---|---|
| Year | 20 | 25 | 30 | 35 |
| 1972 | 0.4014 | 1.3392 | 2.2673 | 2.9440 |
| 1973 | 0.3943 | 1.3201 | 2.2091 | 2.8562 |
| 1974 | 0.3800 | 1.2449 | 2.0990 | 2.7147 |
| 1975 | 0.3622 | 1.1585 | 1.9317 | 2.4954 |
| 1976 | 0.3393 | 1.0928 | 1.8135 | 2.3324 |
| 1977 | 0.3164 | 1.0209 | 1.6733 | 2.1466 |
| 1978 | 0.3165 | 1.0147 | 1.6493 | 2.1019 |
| 1979 | 0.3052 | 1.0052 | 1.6381 | 2.0856 |
| 1980 | 0.3107 | 1.0217 | 1.6831 | 2.1322 |
| 1981 | 0.3154 | 1.0574 | 1.7520 | 2.2214 |
| 1982 | 0.3016 | 1.0193 | 1.7247 | 2.2024 |
| 1983 | 0.2792 | 0.9592 | 1.6188 | 2.0682 |
| 1984 | 0.2932 | 0.9946 | 1.6571 | 2.1087 |
| 1985 | 0.2750 | 0.9479 | 1.6116 | 2.0550 |
| 1986 | 0.2850 | 0.9790 | 1.6600 | 2.1172 |
| 1987 | 0.2851 | 0.9878 | 1.6742 | 2.1401 |
| 1988 | 0.3013 | 1.0257 | 1.7397 | 2.2212 |
| 1989 | 0.3180 | 1.0367 | 1.7528 | 2.2418 |
| 1990 | 0.3243 | 1.0275 | 1.7369 | 2.2274 |
| 1991 | 0.3169 | 0.9701 | 1.6412 | 2.1183 |

The estimates of Total Fertility (TF) associated with cumulated fertility to age $x$ are given by Equation 3. Thus, for example, the estimated Total Fertility implied by the cumulated fertility to age 25 in 1972 is

$$\frac{1.3392}{\exp\left(-\exp\left(-0.0049-1.1367Y^s(x)\right)\right)}=3.5205\,.$$

Averaging the three estimates gives a final estimate of the implied total fertility in each year. Using these estimates in combination with the series of estimates of $\alpha$ and $\beta$, one can derive a series of single-year age-specific fertility rates for each calendar year, again using the relational Gompertz model.

Fertility rates by single years of age for each calendar year are derived using Equation 4. As the matrix produced is large (20 years and 25 ages), only an extract from it is shown in Table 14.11. As an example, the age-specific fertility rate between 16 and 17 in 1990 is given by

$$\begin{aligned}
&f(16,1990)\\
&=TF(16,1990)\cdot\begin{pmatrix}\exp\left(-\exp\left(-\alpha(1990)-\beta(1990)Y^s(17)\right)\right)\\ -\exp\left(-\exp\left(-\alpha(1990)-\beta(1990)Y^s(16)\right)\right)\end{pmatrix}\\
&=2.6516\cdot\begin{pmatrix}\exp\left(-\exp\left(-0.0388-1.1404Y^s(17)\right)\right)\\ -\exp\left(-\exp\left(0.0388-1.1404Y^s(16)\right)\right)\end{pmatrix}\\
&=0.0348.
\end{aligned}$$

Finally, the values of $E(i)$ are derived by applying Equation 5 to the fertility rates in Table 14.11. $E(1)$, the parity equivalent in the 15–19 age group is therefore calculated from

$$\begin{aligned}
E(1)&=\sum_{j=0}^{8}\sum_{m=14}^{18}f(m-j,s-j)\\
&=f(14,1991)+f(15,1991)\\
&\quad+\ldots+f(18,1991)+f(13,1990)\\
&\quad+\ldots f(17,1990)+\ldots+f(6,1983)+\ldots f(10,1983)\\
&=0.07394.
\end{aligned}$$

*TABLE 14.8 Ratios of cumulated fertility to age x and their gompits, by year, Chile*

| | Ratios | | | Gompits $z(x)$ | | |
|---|---|---|---|---|---|---|
| Year | 20 | 25 | 30 | 20 | 25 | 30 |
| 1972 | 0.2997 | 0.5907 | 0.7701 | –0.1864 | 0.6415 | 1.3425 |
| 1973 | 0.2987 | 0.5976 | 0.7734 | –0.1893 | 0.6639 | 1.3590 |
| 1974 | 0.3053 | 0.5931 | 0.7732 | –0.1711 | 0.6493 | 1.3577 |
| 1975 | 0.3126 | 0.5997 | 0.7741 | –0.1508 | 0.6709 | 1.3622 |
| 1976 | 0.3105 | 0.6026 | 0.7775 | –0.1566 | 0.6802 | 1.3798 |
| 1977 | 0.3099 | 0.6101 | 0.7795 | –0.1583 | 0.7050 | 1.3900 |
| 1978 | 0.3119 | 0.6153 | 0.7847 | –0.1529 | 0.7222 | 1.4167 |
| 1979 | 0.3036 | 0.6136 | 0.7855 | –0.1756 | 0.7167 | 1.4209 |
| 1980 | 0.3041 | 0.6070 | 0.7894 | –0.1742 | 0.6948 | 1.4417 |
| 1981 | 0.2983 | 0.6035 | 0.7887 | –0.1904 | 0.6833 | 1.4381 |
| 1982 | 0.2959 | 0.5910 | 0.7831 | –0.1971 | 0.6427 | 1.4085 |
| 1983 | 0.2911 | 0.5925 | 0.7827 | –0.2104 | 0.6475 | 1.4065 |
| 1984 | 0.2948 | 0.6002 | 0.7858 | –0.2001 | 0.6724 | 1.4228 |
| 1985 | 0.2901 | 0.5882 | 0.7842 | –0.2132 | 0.6336 | 1.4145 |
| 1986 | 0.2911 | 0.5898 | 0.7840 | –0.2102 | 0.6386 | 1.4135 |
| 1987 | 0.2886 | 0.5900 | 0.7823 | –0.2173 | 0.6393 | 1.4044 |
| 1988 | 0.2938 | 0.5896 | 0.7832 | –0.2029 | 0.6381 | 1.4092 |
| 1989 | 0.3067 | 0.5915 | 0.7819 | –0.1670 | 0.6441 | 1.4022 |
| 1990 | 0.3156 | 0.5916 | 0.7798 | –0.1425 | 0.6444 | 1.3914 |
| 1991 | 0.3267 | 0.5911 | 0.7748 | –0.1122 | 0.6429 | 1.3658 |

*TABLE 14.9 Values of e(x) and g(x) from the modified Zaba standard (no age shift), selected ages*

| Age $x$ | $e(x)$ | $g(x)$ |
|---|---|---|
| 20 | 1.3539 | –1.3753 |
| 25 | 1.4127 | –0.6748 |
| 30 | 1.2750 | 0.0393 |

Although fertility at ages below 10 can safely be assumed to be zero, these ages are included in the formula for ease of presenting the summations.

***Step 5: Estimate the completeness of birth registration***

For each cohort (defined by age at the end of the final year), the completeness of birth registration is estimated as the ratio of the parity equivalent, $E(i)$, calculated from registered births to reported average parity, $P(i)$, as obtained from the final census (shown in Table 14.1). The results are presented in Table 14.12.

The completeness estimate for the cohort aged 15–19 at the end of 1991 is 0.07/0.14 = 0.5281, or 52 per cent complete. The estimates of completeness of registration of births for women aged 25–29 and 30–34 are both higher and more consistent with each other, suggesting an average level of completeness of around 94 per cent. The estimate for women aged 15–19 is much lower and that for women aged 20–24 suggests that completeness is of the order of 80 per cent. It seems likely that registration of births is substantially less complete among very young mothers while the cumulated completeness of mothers aged 20–24 is also lower because births that they experienced as teenagers are a substantial fraction of their total births.

The estimates of completeness based on the reports of women aged 25–29 and 30–34 are affected relatively little by the excess omission at early ages, so that in this case the best estimate of the completeness of birth registration is probably the average of 0.9382 and 0.9515. The final estimate is therefore 0.9449. Hence, an improved estimate of fertility for 1991 could be obtained by inflating the registered age-specific fertility rates for that year by a factor of 1/0.9449,

*TABLE 14.10 Alpha and Beta parameters of a relational Gompertz model fitted to ages 20, 25 and 30, by year, Chile*

| | | | TF based on cumulant to age | | | |
|---|---|---|---|---|---|---|
| Year | Alpha | Beta | 25 | 30 | 35 | Average |
| 1972 | 0.0049 | 1.1367 | 3.5205 | 3.5525 | 3.5434 | 3.5388 |
| 1973 | 0.0242 | 1.1504 | 3.4057 | 3.4173 | 3.4130 | 3.4120 |
| 1974 | 0.0177 | 1.1367 | 3.2326 | 3.2701 | 3.2598 | 3.2542 |
| 1975 | 0.0270 | 1.1255 | 2.9825 | 3.0074 | 3.0004 | 2.9968 |
| 1976 | 0.0430 | 1.1420 | 2.7704 | 2.7897 | 2.7840 | 2.7814 |
| 1977 | 0.0585 | 1.1503 | 2.5509 | 2.5508 | 2.5498 | 2.5505 |
| 1978 | 0.0828 | 1.1653 | 2.4796 | 2.4782 | 2.4773 | 2.4784 |
| 1979 | 0.0844 | 1.1843 | 2.4518 | 2.4466 | 2.4462 | 2.4482 |
| 1980 | 0.0912 | 1.1982 | 2.4761 | 2.4973 | 2.4902 | 2.4879 |
| 1981 | 0.0850 | 1.2071 | 2.5762 | 2.5995 | 2.5916 | 2.5891 |
| 1982 | 0.0516 | 1.1910 | 2.5607 | 2.6054 | 2.5928 | 2.5863 |
| 1983 | 0.0520 | 1.1990 | 2.4083 | 2.4396 | 2.4300 | 2.4260 |
| 1984 | 0.0712 | 1.2032 | 2.4537 | 2.4750 | 2.4677 | 2.4654 |
| 1985 | 0.0527 | 1.2067 | 2.3781 | 2.4227 | 2.4099 | 2.4035 |
| 1986 | 0.0537 | 1.2038 | 2.4538 | 2.4963 | 2.4840 | 2.4781 |
| 1987 | 0.0479 | 1.2024 | 2.4892 | 2.5248 | 2.5141 | 2.5094 |
| 1988 | 0.0506 | 1.1956 | 2.5788 | 2.6257 | 2.6124 | 2.6057 |
| 1989 | 0.0472 | 1.1654 | 2.6166 | 2.6728 | 2.6579 | 2.6491 |
| 1990 | 0.0388 | 1.1404 | 2.6152 | 2.6780 | 2.6617 | 2.6516 |
| 1991 | 0.0179 | 1.1009 | 2.5207 | 2.5856 | 2.5689 | 2.5584 |

*TABLE 14.11 Age-specific fertility rates by single-years of age and year, Chile*

| | Age | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Year | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 1982 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0012 | 0.0081 | 0.0290 | 0.0574 | 0.0883 |
| 1983 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0010 | 0.0072 | 0.0263 | 0.0530 | 0.0822 |
| 1984 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0011 | 0.0078 | 0.0282 | 0.0562 | 0.0864 |
| 1985 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0009 | 0.0067 | 0.0253 | 0.0517 | 0.0809 |
| 1986 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0010 | 0.0071 | 0.0265 | 0.0538 | 0.0839 |
| 1987 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0010 | 0.0071 | 0.0264 | 0.0538 | 0.0841 |
| 1988 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0011 | 0.0079 | 0.0285 | 0.0571 | 0.0884 |
| 1989 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0016 | 0.0099 | 0.0325 | 0.0615 | 0.0919 |
| 1990 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0021 | 0.0113 | 0.0348 | 0.0634 | 0.0927 |
| 1991 | 0.0000 | 0.0000 | 0.0000 | 0.0002 | 0.0027 | 0.0129 | 0.0363 | 0.0627 | 0.0890 |

or about 5.83 per cent. It should be noted, however, that the adjusted fertility schedule might not be a good indicator of the age pattern of childbearing because of the apparent relatively higher omission of births by young women.

The results presented in Table 14.12 do not suggest that birth registration completeness has been improving over time, so the adjustment factor of 1.0583 can also be applied to the observed age-specific fertility rates for years preceding 1991. However, because the analysis has been truncated at age 30–34 the contributions of fertility rates registered before 1982 to cumulated cohort fertility are small. The estimated adjustment factor thus cannot be validly applied to the fertility rates registered before that date. Although not observed in the case of Chile, where there is evidence of a trend towards more complete registration, such as a tendency for the estimates of completeness to decline with age, registered births for particular years should not be adjusted using a scaling factor derived from an average over many years.

However, none of these considerations explains adequately the low estimates of completeness obtained for the cohorts aged 15–19 and, to a lesser extent, those aged 20–24. Of course, the procedure used to split the period fertility rates is not perfect and it is most likely to be inaccurate at 15–19, but possible methodological inaccuracy cannot explain the large differential observed. Furthermore, since average parities were calculated without making any adjustment for non-response, they are more likely to be too small than too large. Hence, on the basis of this evidence alone, it would appear that birth registration in Chile is less complete for young mothers than for older women.

*TABLE 14.12 Estimates of completeness of birth registration, Chile 1991*

| | | Age group | | | |
|---|---|---|---|---|---|
| | | 15–19 | 20–24 | 25–29 | 30–34 |
| MCEB from last census (applies to end year) | 1991 | 0.14 | 0.69 | 1.37 | 2.02 |
| Parity equivalents (applies to end year) | 1991 | 0.07 | 0.57 | 1.29 | 1.92 |
| Completeness | | 0.5281 | 0.8207 | 0.9382 | 0.9515 |
| Average completeness | | | | 0.9449 | |

## FURTHER VARIANTS

An extension of the method, using parities from two surveys five or ten years apart, combined with vital registration data covering births over a 20-year period, is described on pages 55ff of *Manual X* (UN Population Division 1983). The main assumptions made in the version described above to adjust period fertility on the basis of cumulated cohort fertility from the beginning of childbearing – namely, that registration completeness be constant both by age of mother and by period – are no longer required. However, parity increments are very sensitive to changes in the completeness of reporting of children ever born, as a result of which the estimates of registration completeness obtained by the method described on pages 55ff of *Manual X* are also quite sensitive to such changes, which are generally most marked for older women.

The main differences between the procedure described on pages 55ff of *Manual X* and that described above are that all calculations in the variant approach are limited to the inter-survey period, and all cohorts of reproductive age at the time of the second survey are considered. In doing so, the method compares the increments in fertility according to the registration system over the inter-survey period, rather than completely cumulated fertility for cohorts of women from the start of their childbearing. There is a similar difference in respect of inter-survey parity increments for cohorts. The measure of completeness is thus based on the ratio of inter-survey cumulated fertility, and inter-survey parity increments.

## FURTHER READING AND REFERENCES

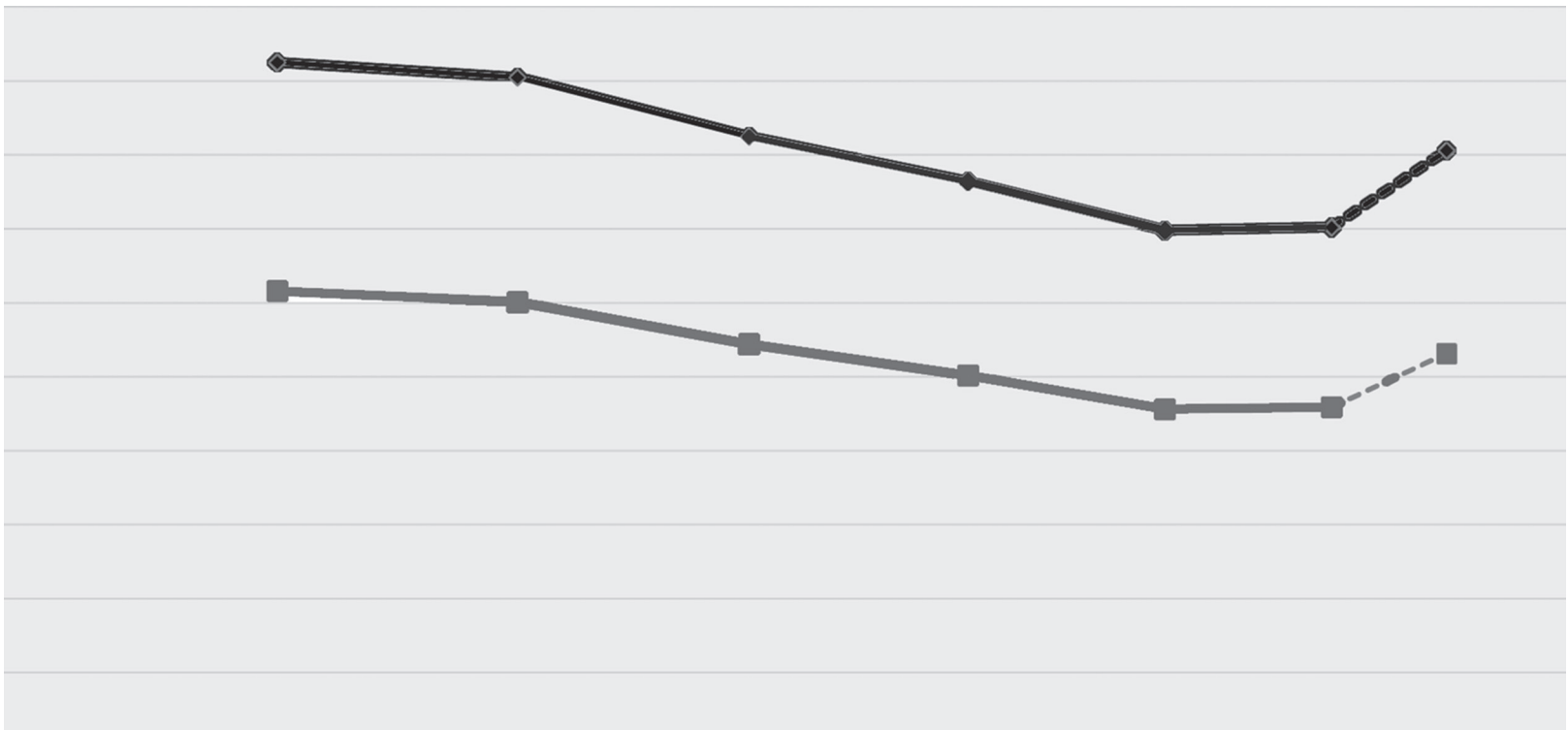
The method was originally set out on pages 45–55 of *Manual X* (UN Population Division 1983). Given its dependence on a long series of vital registration data, the method has not been applied widely. However, as more developing countries seek to implement or improve systems of vital registration, the method might become more important in future as a check on the quality of those data.

The only substantial change made in the implementation of the method as described here is that, whereas the original formulation relied on tabulated coefficients based on the Brass fertility schedule to apportion fertility to non-standard age groups – a long, tedious and error-prone process – the approach adopted here is simpler, making use of the relational Gompertz model to do the same thing. In effect, what is now done is that a separate relational Gompertz model is fitted to the fertility rates observed in the vital registration data in each year. The values of the parameters alpha and beta so derived can then produce estimates of the implied total fertility based on the observed fertility to each of ages 25, 30 and 35. The average of these three estimates of total fertility is then taken as the estimate of total fertility in each year, thereby permitting the calculation of age-specific fertility rates by single years of age in each calendar year. Using these, the calculation of cumulated cohort fertility is simple, as the relevant fertility rates in each year simply need to be summed.

## REFERENCES

UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

# Child Mortality



# Chapter 15 Introduction to child mortality analysis

*Kenneth Hill*

The level of child mortality in a country has long been treated as an index of general development. Over the last two decades, the international community has established periodic targets for the reduction of child mortality. The World Summit for Children in 1990 set the goal of reducing ${}_5q_0$ by one-third, or to below 70 per thousand, between 1990 and 2000. The 1994 International Conference on Population and Development reiterated this goal, but set the additional longer-term target of reducing ${}_5q_0$ to below 45 per thousand by 2015. The Millennium Summit in 2000 then adopted a set of eight development goals, of which the fourth included a target of reducing ${}_5q_0$ by two-thirds between 1990 and 2015. The report of the Commission on Information and Accountability for Women's and Children's Health, established by the Secretary-General of the United Nations, has recently reaffirmed the importance of frequent reporting on ${}_5q_0$. It is thus clear that the measurement of infant and child mortality has a very high priority both at the national and at the international level. This introduction discusses broad options for such measurement, with details of the various methods discussed in other chapters.

## DATA FOR ESTIMATING UNDER-5 MORTALITY

Data on child mortality from sources other than civil registration systems (and often the data in civil registration systems as well) come almost exclusively from mothers. The data are collected either about all (generally live-born) children or about a specified child or group of children. They are collected with varying amounts of detail, bearing in mind the difficulty of recalling details of births long in the past. Lack of detail in data collection may subsequently be compensated for at the analysis stage by making assumptions, primarily about past age patterns of child mortality and fertility, but with increased uncertainty.

The fact that data are collected from mothers is important, because it raises the issue of selection bias. If all the children (alive or dead) in a target population over an extended period had mothers who were still alive and also still in the population, then each child has an equal probability of being reported and there will be no bias. However, in practice mothers are themselves subject to mortality and migration risks and, as a result, at a subsequent date it will no longer be the case that all children in a target population and only those children in the target population will be reported. The largest threat in terms of bias is in populations substantially affected by HIV/AIDS. In such populations, in the absence of extensive use of antiretrovirals, there will be a strong correlation between the mortality risks of mothers and of their children, such that child mortality as reported by surviving women will underestimate true child mortality. This issue is discussed in more detail in Effects of HIV on methods of child mortality estimation. The longer the time lag between the birth of a child and the data collection, the greater the potential for such biases.

We distinguish below between data collected through a household interview (including data collection exercises that aim for universal coverage of a population, such as a census, and those that collect data from a representative sample of a population, such as Demographic and Health Surveys) on the one hand and data collected at a health facility, such as a birthing centre, on the other.

## DATA COLLECTED BY HOUSEHOLD INTERVIEWS

Surveys collecting data through household interviews can range from universal (a population census) to small and focused samples. The amount and complexity of data that can be collected in an interview varies by survey size: in general, large surveys permit only short questionnaires and limited interviewer training, whereas small surveys can (subject to financial constraints) use much longer questionnaires and interviewers can be trained more carefully to collect much

more detailed information. More detailed information can support analysis of a wider range of indicators, whereas simple data for a larger sample or for the whole population permit much greater disaggregation by geography or socio-economic factors. The two sources may thus be seen as complementary rather than competing. In this section, we describe types of data from the most complex to the simplest. We then proceed to describe the process of data evaluation and analysis.

**The full birth history**

A full birth history (FBH) collects information from surveyed mothers about each live-born child to whom the woman has ever given birth. In most settings, information is collected from all women of reproductive age (normally 15 to 49). However, in situations where it is considered inappropriate to ask never-married women about childbearing, questions may be limited to ever-married women. The minimum information collected about each child is name, date of birth (often collected as month and year), whether the child is still alive, and if the child has died, age at death. In Demographic and Health Surveys (DHS) age at death is collected in days if the child dies before 28 days, in months if the child dies between 28 days and 24 months, and in years thereafter. From an analyst's point of view the use of years as units is less than optimal since deaths cannot be located with adequate precision in calendar time.

In the DHS, data are collected by order of the birth from the earliest to the latest, with probes between each to make sure no child is unreported. Additional information is often collected on whether the child was one of a multiple birth, the child's sex, whether the child is living with the mother, and, if so, the line number of the child on the household listing. The data collected thus allow the location of each birth and, if the child has died, its death in time, permitting the calculation of conventional occurrence-exposure rates and use of standard life table analytic methods to estimate mortality risks for age ranges of childhood.

In the DHS, the full birth history is prefaced by a detailed summary birth history (SBH), whereby each woman is asked about the number of sons and daughters living at home, the number of sons and daughters living elsewhere, and the number of sons and daughters that have died. The FBH is then collected, and the FBH and SBH are field-edited to ensure consistency. Because this edit is done in the field, DHS provides no evidence on the frequency, magnitude or direction of discrepancies between the FBH and the initial SBH.

Some surveys have collected full pregnancy histories, adding all pregnancies that did not result in a live birth. This has not been the norm in the DHS, although the model questionnaire does include the collection of data about pregnancy losses that occurred in the five years or so before the survey. It has been argued that a pregnancy history will collect better data about live births than a birth history, by increasing the likelihood that very early neonatal deaths will get reported, but there are no experimental data to bear this out.

As will be clear, the FBH is a complex data collection instrument that requires extensive fieldworker training and careful field supervision. It is only used when the woman herself can be interviewed; proxy reports are not allowed. It is not appropriate for very large-scale data collection exercises such as a population census.

**The truncated birth history**

Some surveys have used truncated birth histories (TBH) to reduce fieldwork costs and possible interviewee fatigue. A TBH collects the same information as a FBH about each birth, but limits the number of births for which information is collected either by time period (for example, to births in the preceding 5 years) or by number (for example, for the most recent births up to a limit of three). Experience suggests that TBHs are a risky form of data collection for both theoretical and practical reasons. A time limitation drastically reduces information about mortality risks of somewhat older children (for example with a five-year cut-off, only children born between 4 and 5 years ago will provide information about mortality between the ages of 4 and 5, and will provide no information about mortality above age 5). It also runs a practical risk that children that have died may be intentionally shifted out of the time window. A number limitation raises concerns about selection bias (women with more than 3 recent births will only report on the most recent 3) and also selective reporting (potentially favouring reporting on surviving children rather than children who have died).

**The summary birth history**

In a summary birth history (SBH), the only information collected is aggregate numbers of children ever born and children still alive (or dead). The amount of detail varies, from just two questions (number of children ever born and number of children still alive) to the detailed SBH

asking about boys and girls separately, and enquiring about surviving children living at home versus those living elsewhere, as described above for the DHS FBH sequence of questions. It is believed, without any empirical basis as far as we know, that the more detailed form of the questions improves data quality. Unlike the FBH, the SBH is often used with proxy reports. In a census, for example, a single household informant will usually provide the information. The informant will sometimes be the woman herself, but by no means always. In some cases a male head of household will provide information for all the women living there.

The SBH does not provide a basis for direct calculation of standard mortality indicators. Events (deaths) are not located explicitly in time, and there is no basis for calculating precise exposure times. Thus conventional occurrence-exposure rates cannot be calculated from the data. Instead, the data are analysed using an indirect estimation of child mortality approach, whereby the age (or sometimes the duration of marriage or the time since first birth) of a group of women is taken as an indicator of the average exposure time of their children to the risk of dying, and model age patterns of fertility and mortality are used to convert the proportions dead of children ever borne by women in a group into a standard life table function.

**Survival of a recent birth**

The SBH does not in practice provide a very recent estimate of child mortality, for reasons explained in Chapter 16. Censuses and large-scale surveys have attempted to obtain a more recent estimate by asking about the survival or otherwise of a recent child. (Information on a recent birth is also used as a basis for fertility estimation.) There are essentially two forms of the questions in common use. One asks whether a woman had a live birth in a specified time period (such as the 12 months before the survey) and whether that child is still alive. The other asks for the date (month and year) of the woman's most recent live birth and whether the child is still alive. Either form of question can be used to calculate the proportion alive of (a large majority of) children born in the 12 months before the survey. If births are distributed evenly across the 12 months, the proportion alive estimates the life table parameter $1-\left(\frac{{}_1L_0}{l_0}\right)$.

The second form of question can also be used to calculate survival to later ages of childhood, as proposed by the Brass-Blacker procedure below.

In practice, as described in Chapter 4, the number of births reported in the last 12 months is often implausible (usually too low), raising issues about selectivity or reporting bias possibly related to a child's survival. Conversion of the parameter $1-\left(\frac{{}_1L_0}{l_0}\right)$ into the generally-used index ${}_5q_0$ is also dependent on the choice of a model life table system.

Blacker and Brass (2005) suggest using data from the second format (date of most recent birth) to estimate child mortality from survival of most recent births in the two years before a census or survey. The authors show that for typical age patterns of child mortality, the proportion dead among such births can in the absence of selection bias be converted into an estimate of the infant mortality rate ${}_1q_0$ simply by multiplying by a factor of 1.09. However, this adjustment factor is based on the assumption that the most recent births in the last 24 months are not selected from all births in the last 24 months (i.e. including all births of women with more than one birth in the period) on the basis of survival. This assumption turns out to be surprisingly far off the mark. Analysis of DHS full birth history data shows that for women with more than one birth in the period, the earlier birth(s) are much less likely to have survived than the most recent. A major reason for this is that birth intervals after an early child death are systematically shorter than those after a child that survives infancy, with the consequence that births in the last 24 months that do not survive are more likely to be followed by a subsequent (potentially surviving) birth. When this happens, the child that died does not get reported as a most recent birth. In a sample of African DHSs, the proportion dead of most recent births in the last 24 months is biased downwards relative to the proportion dead among all births in the 24 months by over 20 percent on average. Even in Bangladesh, a population with long birth intervals where the effect would be expected to be smaller, the bias still reaches almost 16 percent.

In view of concerns about both data quality in the case of survival of most recent births in the 12 months before a survey and selection bias when data for 24 months are used, the analysis of information on survival of most recent births is not recommended.

## DATA COLLECTED AT HEALTH FACILITIES

A major expense of household surveys is the cost of getting an interviewer to the (correct) household. Much of this expense can be eliminated by taking advantage of respondents

coming to the interviewer, such as on visits to health facilities. Health facilities are also likely to record births and deaths that occur at the facility as part of a routine health management information system. The problem with using such data for child mortality estimation is selection bias in that we can never be sure that the women who visit a facility are representative of all mothers. To improve coverage, experiments are being conducted to find out whether health extension workers or other health staff working in communities can collect adequate data on births and deaths. Such an approach is akin to a sample registration system.

Two approaches to collecting data from women at the health facility have been proposed, and one, the preceding birth technique (PBT), has been extensively developed (Brass and Macrae 1984). In essence, what this method does is ask women coming to a facility to give birth whether their previous birth, if they had one, is still alive. Given typical birth interval distributions, the proportion of previous births that have died can be interpreted as a probability of dying by an exact age of childhood. The data are extremely easy and inexpensive to collect, and can provide an indication of child mortality trends even if the level may be affected by selection bias. Such trends, however, will not be accurate if the selection bias changes materially over time.

The second approach starts with the assumption that women who visit a health facility are a biased sample of all mothers, but does not assume *a priori* the direction of the bias. Instead of assuming that women with healthier than average children or women with less healthy children will predominate, it aims to estimate the selection probabilities so as to be able to adjust for bias. For example, women visiting a health facility could be asked their age, children ever born and children still alive, plus a number of additional questions about their socio-economic characteristics. The children ever born and children still alive could then be modelled onto the socio-economic structure of the whole population, available for example from a population census. To our knowledge, this approach has never been tested.

## EFFECTS OF HIV ON METHODS OF CHILD MORTALITY ESTIMATION

All estimation methods for child mortality based on mother's reports on the survival of their children are subject to selection biases. Although migration and selective non-response may introduce bias, the greatest threat to child mortality estimates based on reports of women arises from a generalized HIV/AIDS epidemic. Vertical transmission of HIV from mother to child during pregnancy, delivery and through breastfeeding in the first few months of life increases the risk that the child will be HIV-positive (HIV) by as much as 35 per cent in the absence of antiretrovirals (ARVs), and over 60 per cent of HIV+ children will die before their 5th birthday in the absence of treatment with ARVs (Schneider, Zwahlen and Egger 2004; Todd, Glynn, Marston *et al.* 2007). Since the mothers also suffer elevated mortality risks, the deaths of many of these HIV+ children, particularly those born 5 years or more before interview, will not be reported. Overall child mortality will therefore be under-estimated, whether using direct child mortality estimation or indirect child mortality estimation.

### Effects on direct child mortality estimation

Only one known analysis of the magnitude of bias in direct child mortality estimates has been carried out using real data rather than simulations. Hallett, Gregson, Kurwa *et al.* (2010) use data from a prospective open cohort in Manicaland, Zimbabwe to measure the bias introduced by deaths of HIV+ mothers. The cohort was interviewed between July 1998 and February 2000, with follow-up interviews at three and five years. From 1998 to 2005, HIV prevalence in the study population fell from 22 per cent to 18 per cent. In the final round in 2005, a full birth history was collected from surviving women, and U5MR was estimated for the period 1998 to 2005, a seven-year period as opposed to the usual DHS five-year period. The direct estimates were then compared to true values, adding back the child mortality experience of women who had died before 2005. The bias, calculated as the estimates from surviving mothers divided by the estimates for all mothers, was 6.7 per cent for the IMR and 9.8 per cent for U5MR. Hallett, Gregson, Kurwa *et al.* (2010) also developed a model of bias, which they applied to Zimbabwe and six other countries with moderate or high HIV prevalence for the period 1980 to 2015. They did this using UNAIDS prevalence data and DHS estimates for pre-epidemic periods. The model indicates that bias in direct estimates increases with (a) the duration of the epidemic; and (b) the time before survey of the estimate. Conversely, bias in direct estimates decreases with the level of background, non-HIV child mortality.

Walker, Hill and Zhao (2012) developed a simple cohort component projection model, separating births into those to HIV-negative (HIV–) mothers (where the children are

assumed not to be infected), those to HIV+ mothers but where the children are not themselves infected, and those to HIV+ mothers where the children are infected at birth or subsequently. The first two streams are assumed to experience background mortality (from model life tables) whereas the third stream is assumed to experience a probability of dying by age 5 of 62 per cent. HIV+ mothers are then aged forward to the date of a survey allowing for their excess mortality, and the U5MR estimated from reports of surviving women is compared to that which would have been observed had all women survived to the survey. No adjustment is made for prevention or treatment. As with the analysis by Hallett, Gregson, Kurwa *et al.* (2010), the extent of bias depends on the HIV prevalence and its past trajectory, the level of background under-5 mortality and the time period before the survey to which the estimate refers. It is thus not possible to provide a simple way to assess the magnitude of bias. However, as a general guide, some estimates of bias for countries collecting birth histories around the middle of the last decade, before ARV treatments were widely available, are shown in Table 15.1 for time periods 1–5, 6–10 and 11–15 years before each survey. It is important to remember, however, that bias is a function of non-HIV child mortality, which is not easy to estimate, and HIV prevalence, which is usually estimated with error.

Bias is highest for the period 6–10 years before the survey, exceeding 10 per cent if the HIV prevalence has exceeded 5 per cent. It is important to remember that ARV use to prevent mother-to-child transmission and to extend survival times will have a quick effect on reducing bias for the most recent time period, but bias for past time periods will persist for a decade or more after effective therapy is introduced.

**Effects on indirect child mortality estimation**

Few studies have explored the impact of HIV on indirect estimates of child mortality. HIV will affect the accuracy of indirect estimates not only because of the association between mortality of children and that of their mothers but also because of the effects of HIV on age patterns of child mortality, and its implications for approaches that infer fertility patterns from observed parity ratios. Child mortality risks can also no longer be assumed to be independent of the age of the mother. However, in one respect, indirect estimates may be less affected by selection for maternal survival than direct estimates, because the analysis is carried out by age group. Mothers under the age of 25 are unlikely to have died from HIV/AIDS, so reports of child survival based on age groups 15–19 and 20–24, and even 25–29, may be little biased by HIV. However, it is the age groups 15–19 and 20–24 that are most biased by other selection effects, so this may not be a huge help.

Ward and Zaba (2008) estimate likely bias from HIV on indirect child mortality estimates given a stable (constant incidence and mortality effects) epidemic. Their model shows that the bias for estimates based on women under age 30 for adult prevalence of 10 per cent or less will not exceed 5 per cent, and even for prevalence up to 30 per cent will scarcely exceed 10 per cent. These findings are reassuring. Of course, the HIV epidemic has been anything but stable, rising sharply across many countries to around 2000 and declining both in prevalence and impact since then. However, the dynamic of the epidemic will tend to reduce bias below the levels estimated by the Ward and Zaba model.

Mutemaringa (2011) compares indirect estimates derived from DHSs for Zimbabwe, Kenya, Lesotho, Malawi,

| Country | DHS Year | Approx. HIV prevalence 2005 (%) | Assumed background U5MR | Estimated bias by period before survey (%) | | |
|---|---|---|---|---|---|---|
| | | | | 1–5 | 6–10 | 11–15 |
| Cote d'Ivoire | 2005 | 4.6 | 125 | 4.0 | 6.6 | 3.3 |
| Kenya | 2003 | 7.1 | 75 | 8.0 | 14.1 | 6.7 |
| Lesotho | 2004 | 23.4 | 75 | 13.2 | 15.7 | 2.1 |
| Namibia | 2006–07 | 15.3 | 50 | 13.7 | 22.7 | 10.4 |
| Zambia | 2007 | 15.0 | 150 | 6.8 | 13.9 | 13.0 |
| Zimbabwe | 2005 | 18.0 | 75 | 16.6 | 31.4 | 25.6 |

Source: Walker, Hill and Zhao (2012)

*TABLE 15.1 Estimates of bias for estimates of U5MR for periods 1–5, 6–10 and 11–15 years before each survey: selected sub-Saharan African countries*

Namibia and Zambia to direct estimates from the same surveys. The author confirms that the bias primarily arises from the survivorship correlation. The bias in the estimate based on reports of women aged 25–29 is in three cases out of six less than 5 per cent, although in two cases, Zimbabwe and Namibia, the bias exceeds 20 per cent. The bias of estimates based on reports of women aged 30–34 and 35–39 generally exceeds 20 per cent, and in Kenya and Namibia exceeds 30 per cent.

The conclusion we draw from these analyses is that estimates of child mortality derived from reports of women aged 25–29 concerning their children ever borne and surviving will not be greatly affected by even a generalized HIV epidemic. Drawing on the pattern of bias by HIV prevalence found by Ward and Zaba (2008), the child mortality estimate obtained by standard analysis of a summary birth history could be adjusted upwards by three points per thousand for every 10 percentage points of HIV prevalence.

## FURTHER READING AND REFERENCES

In addition to the works cited above, the interested reader is directed to the work of Hill (1984, 1991), and *Manual X* (UN Population Division 1983) for further historical reading on indirect estimation; to Rajaratnam, Tran, Lopez *et al.* (2010) for recent developments in modelling under-5 mortality, as well as Preston (1985) and Rutstein and Rojas (2003) for a description of direct estimation of child mortality from survey data.


Blacker J and W Brass. 2005. "The estimation of infant mortality from proportions dying among births in the past 24 months", *Southern African Journal of Demography* **10**(1–2):25–42.

Brass W and S Macrae. 1984. "Childhood mortality estimated from reports on previous births given by mothers at the time of a maternity: I. Preceding-births technique", *Asian and Pacific Census Forum* **11**(2):5–8. http://hdl.handle.net/10125/3561

Hallett TB, S Gregson, F Kurwa, G Garnett *et al.* 2010. "Measuring and correcting biased child mortality statistics in countries with generalized epidemics of HIV infection", *Bulletin of the World Health Organization* **88**(10):761–788. doi: 10.2471/BLT.09.071779

Hill K. 1984. "An evaluation of indirect methods for estimating mortality," in Vallin, J, JH Pollard and L Heligman (eds). *Methodologies for the Collection and Analysis of Mortality Data*. Liège, Belgium: Ordina Editions, pp. 145–176.

Hill K. 1991. "Approaches to the measurement of childhood mortality: A comparative review", *Population Index* **57**(3):368–382. doi: 10.2307/3643873

Mutemaringa T. 2011. "Impact of HIV on estimates of child mortality derived using the summary birth history (CEB/CS) method." Unpublished MPhil thesis, Cape Town: University of Cape Town.

Preston SH. 1985. "Mortality in childhood: Lessons from the WFS," in Cleland, J and J Hobcraft (eds). *Reproductive Change in Developing Countries*. Oxford: Oxford University Press, pp. 253–272.

Rajaratnam JK, LN Tran, AD Lopez and CJL Murray. 2010. "Measuring under-five mortality: Validation of new low-cost methods", *PLoS Medicine* **7**(4):e1000253. doi: 10.1371/journal.pmed.1000253

Rutstein S and G Rojas. 2003. *Guide to DHS Statistics*. Calverton, MD: ORC Macro.

Schneider M, M Zwahlen and M Egger. 2004. *Natural history and mortality in HIV-positive individuals living in resource-poor settings*. http://www.epidem.org/Publications/unaids%20HQ_03_463871%20final.pdf

Todd J, JR Glynn, M Marston, T Lutalo *et al.* 2007. "Time from HIV seroconversion to death: a collaborative analysis of eight studies in six low- and middle-income countries before highly active antiretroviral therapy", *AIDS* **21**(Suppl 6):555–563. doi: 10.1097/01.aids.0000299411.75269.e8

UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

Walker PN, K Hill and F Zhao. 2012. "Child mortality estimation: Methods used to adjust for bias due to AIDS in estimating trends in under-five mortality", *PLoS Medicine* **9**(8):e1001298. doi: 10.1371/journal.pmed.1001298

Ward P and B Zaba. 2008. "The effect of HIV on the estimation of child mortality using the children surviving/children ever born technique", *Southern African Journal of Demography* **11**(1):39–73.

# One Census Methods

# Chapter 16 Indirect estimation of child mortality

*Kenneth Hill*

## DESCRIPTION OF THE METHOD

Indirect methods, pioneered by Brass and Coale (1968), estimate child mortality from information on aggregate numbers of children ever born and children still alive (or dead) reported by women classified by age group (or alternatively grouped by time since first birth, or marital duration). Such information is described as a summary birth history (SBH). The amount of detail collected varies, from just two questions (number of children ever born and number of children still alive) to the most detailed SBH asking separately about boys and girls, and enquiring separately about surviving children living at home versus those living elsewhere described in Chapter 15 as part of the DHS full birth history sequence of questions. The proportion dead of children born to women by age (or time since first birth, or duration of marriage) reflects the level of child mortality, but is also affected by other things, primarily the age pattern of childbearing and the age pattern of child mortality. Young mothers generally have young children, who have been exposed to the risk of death for short, recent periods; the proportion dead for such mothers thus reflects child mortality risks to an early age. Older mothers, in contrast, have a mix of young and older children exposed to the risk of dying for longer periods on average further in the past. Through models of fertility and child mortality, the proportions dead are converted into probabilities of dying by exact ages of childhood, ${}_nq_0$. The older the women, the greater the value of $n$.

If mortality has changed over time, the estimated probabilities of dying reflect the mortality rates that have prevailed at a range of ages and dates. Fortunately, a 'time location' method has been developed that estimates how many years previously each proportion dead approximates period probabilities of dying. These intervals increase with the age of respondents. Thus, if the probabilities of dying estimated from the reports of different age groups of woman are translated into a common index of mortality, these statistics will refer to different dates and can be used to infer the broad trend in mortality over time.

## DATA REQUIREMENTS AND ASSUMPTIONS

### Tabulations of data required

- Number of women, grouped by five-year age, duration of marriage, or time since first birth.
- Number of children ever born alive by women by relevant (age, time since first birth, or marriage duration) five-year group.
- Number of children born alive by the women that have died before (or are still alive at) the time of the survey, by relevant five-year group.
- Number of births in the year before the survey by five-year age group (optional).

### Important assumptions

- Population age patterns of fertility and child mortality are adequately represented by the model patterns used in developing the method.
- In any time period, mortality of children does not vary by five-year grouping of mothers.
- No correlation exists between mortality risks of children and survival of mothers (by mortality or migration) in the population (see Effects of HIV on methods of child mortality estimation, in Chapter 15).
- Any changes in child mortality in the recent past have been gradual and unidirectional.
- Cross-sectional average numbers of children ever born by age (or by duration of marriage or time since first birth) adequately reflect the appropriately-defined cohort patterns of childbearing.

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Data quality assessment for an SBH involves a subset of analyses described for the analysis of a full birth history (FBH) in Chapter 17. Since an SBH contains no information on dates of individual births, assessment is limited to analysis of aggregates, tabulated by age group of mother (or duration of marriage or time since first birth, if such information is available); single-year tabulations can also be revealing if there is substantial age heaping. Once again, assessments are carried out of internal plausibility within the data set itself, and of external consistency with other data sets for the population. Of internal assessments, a first check should be of the average number of children ever born (CEB) in each group of women. (Note that the appropriate denominator for these calculations is all women, not the number of mothers or number of ever-married women). Unless fertility has been rising, the average number of CEB should increase with five-year group. A second check should be of the average number of children dead (CD) by each five-year group. Unless child mortality or fertility has been increasing, the average number of CD should also increase with age. If information is available by sex of children ever born, sex ratios of births should be calculated. As with the FBH, any tendency for such ratios to deviate far from 100 to 106 males per 100 females, or to rise with age (or marital duration or time since first birth) should be interpreted as warning signs unless the population is known to practice sex-selective abortion. Among external checks, the cohort comparisons of CEB and CD described for the FBH in Direct estimation of child mortality are often revealing.

## CAVEATS AND WARNINGS

- Care is needed with data editing. Women with missing data on numbers of children ever borne, or numbers of children dead (or surviving), or both, should be excluded from the analysis. However, women with no children must be included in the analysis.
- Care is also needed with imputation. Information on CEB in large surveys is often differentially missing for childless women (see Chapter 4, but note that for this purpose the correction is not needed because it would affect each average parity by the same proportional amount). Hot-deck imputation may cause serious bias.
- The assumption that mortality of children does not vary by five-year grouping of mothers is generally incorrect when the time dimension is age. Children of young mothers appear to have systematically higher mortality than children born to women after age 25. As a result, indirect estimates derived from women aged 15–19 (particularly) and 20–24 (to some extent) tend to overestimate the population-level child mortality. Corrections for these effects are not widely used (Collumbien and Sloggett 2001). It is in part this distortion that has led to the development of methods based on duration of marriage or time since first birth. However, these variants have their own limitations, as described later.
- Application of the method in populations with generalized HIV epidemics requires great care (see Effects of HIV on methods for child mortality estimation in Chapter 15).

## APPLICATION OF METHOD

### Summary birth history by five-year age group/time since first birth/duration of marriage of mother

#### *Step 1: Calculate proportions dead of children ever born, ${}_5PD_x$*

For each five-year group $(x,x+5)$ of women, the proportions of children dead are calculated by dividing the number of children dead by the number ever born.

#### *Step 2: Calculate average numbers of children ever born to women in each five-year group, ${}_5P_x$*

For each five-year group of women, divide the reported number of children born ${}_5CEB_x$ by the number of women ${}_5N_x$ in the group. Note that if the classifying variable being used is age, the denominator should be all women, regardless of marital status or childbearing history.

#### *Step 3: Select a model life table family*

The age pattern of child mortality has an important bearing on the translation of a proportion dead into a standard ${}_nq_0$, and on the translation of that ${}_nq_0$ into a common index such as U5MR. In a population for which fairly recent FBH data exist, the model can be selected on the basis of the FBH by plotting estimates of ${}_4q_1$ against those of ${}_1q_0$ on a graph showing the corresponding relationships in Coale-Demeny and United Nations model life tables. If no appropriate FBH estimates exist, the selection of a model can be based on child mortality patterns in a neighbouring population. A perfect fit of the data to any model is unlikely. The analyst should select the model that best represents the range of observations available. The same model family should

be used in Steps 4, 5 and 6. Steps 4 and 5 have different variants, depending on whether the data being analysed are classified by age of mother (stream a), time since first birth of mother (stream b), or duration of marriage of mother (stream c).

**Data classified by age of mother (stream a)**

***Step 4a(1): Estimate the mean age of the age-specific fertility schedule***

This sub-step is only required for analysis using one of the United Nations model life tables for developing countries. It is calculated from age-specific fertility rates ${}_5f_x$ as follows:

$$\bar{m} = \frac{\sum_{x=15,5}^{45} {}_5f_x.(x+2)}{\sum_{x=15,5}^{45} {}_5f_x}.$$

The $(x+2)$ term in the numerator represents the mid-point of the age group $x$ to $x+5$ at the time when the births occurred. This assumes that age-specific fertility rates are calculated from information on births in the year before the survey classified by age of woman at the time of the survey (see Chapter 5 for more details). If age-specific rates are calculated from registered births by age of mother at the birth, the term would be $(x+2.5)$. Note that it is only the relative pattern of fertility by age that determines the mean, so there is no need to adjust level by, for example, a relational Gompertz model, before calculating it.

***Step 4a(2): Estimate ${}_nq_0$ from each ${}_5PD_x$***

Once a model life table family $j$ has been identified, appropriate parameters $a(x,j)$, $b(x,j)$ and $c(x,j)$ (and $d(x,j)$ if a UN model life table is being used) are substituted into the following equation:

$$\frac{{}_nq_0}{{}_5PD_x} = a(x,j) + b(x,j)\times\frac{{}_5P_{15}}{{}_5P_{20}} + c(x,j)\times\frac{{}_5P_{20}}{{}_5P_{25}} + d(x,j)\times\bar{m}.$$

Note that $d(x,j)$ is zero unless using the UN model life tables.

For each age group $(x,x+5)$, ${}_nq_0$ is estimated by multiplying the right-hand side of the equation by the empirically observed ${}_5PD_x$ (Table 16.1).

***Step 5a: Estimate the time reference t(x) of each estimated ${}_nq_0$***

Using the model life table family $j$ identified as being appropriate, parameters $e(x,j)$, $f(x,j)$ and $g(x,j)$ are substituted into the following equation:

$$t(x) = e(x,j) + f(x,j)\times\frac{{}_5P_{15}}{{}_5P_{20}} + f(x,j)\times\frac{{}_5P_{20}}{{}_5P_{25}}.$$

The location of the estimates in calendar time is readily achieved by subtracting the $t(x)$ from the census or survey date (Table 16.2).

**Data classified by time since first birth of mother (stream b)**

***Step 4b: Estimate ${}_nq_0$ from each ${}_5PD_x$***

Once a model life table family $j$ has been identified, appropriate parameters $a(x,j)$, $b(x,j)$ and $c(x,j)$ are substituted into the following equation:

$$\frac{{}_nq_0}{{}_5PD_x} = a(x,j) + b(x,j)\times\frac{{}_5P_{15}}{{}_5P_{20}} + c(x,j)\times\frac{{}_5P_{20}}{{}_5P_{25}}.$$

Note: the coefficients and values of $n$ in ${}_nq_0$ have been updated by Hill from those published in Hill and Figueroa (2001).

For each age group $(x,x+5)$, ${}_nq_0$ is estimated by multiplying the right-hand side of the equation by the empirically observed ${}_5PD_x$ (Table 16.3).

***Step 5b: Estimate the time reference t(x) of each estimated ${}_nq_0$***

Using the model life table family $j$ identified as being appropriate, parameters $e(x,j)$, $f(x,j)$ and $g(x,j)$ are substituted into the following equation:

$$t(x) = e(x,j) + f(x,j)\times\frac{{}_5P_{15}}{{}_5P_{20}} + g(x,j)\times\frac{{}_5P_{20}}{{}_5P_{25}}.$$

The results are presented in Table 16.4. Note: the coefficients and values of $n$ in ${}_nq_0$ have been updated by Hill from those published in Hill and Figueroa (2001).

**Data classified by duration of marriage of mother (stream c)**

***Step 4c: Estimate ${}_nq_0$ from each ${}_5PD_x$***

Once a model life table family $j$ has been identified, appropriate parameters $a(x,j)$, $b(x,j)$ and $c(x,j)$ are substituted into the following equation:

$$\frac{{}_n q_0}{{}_5 PD_x} = a(x,j) + b(x,j) \times \frac{{}_5 P_{15}}{{}_5 P_{20}} + c(x,j) \times \frac{{}_5 P_{20}}{{}_5 P_{25}} \cdot$$

For each duration of marriage group $(x,x+5)$, ${}_n q_0$ is estimated by multiplying the right-hand side of the equation by the empirically-observed ${}_5 PD_x$ (Table 16.5).

***TABLE 16.1*** *Values of a(x,j), b(x,j), c(x,j) and optional d(x,j) for estimating probabilities of dying by exact ages of childhood from proportions dead of children ever born classified by age of mother*

| | | Age group of mother and value of $n$ in ${}_n q_0$ | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Family *j*** | | **15–19** | **20–24** | **25–29** | **30–34** | **35–39** | **40–44** | **45–49** |
| | | **1** | **2** | **3** | **5** | **10** | **15** | **20** |
| Princeton 'North' | *a*(*x*,*j*) | 1.1119 | 1.2390 | 1.1884 | 1.2046 | 1.2586 | 1.2240 | 1.1772 |
| | *b*(*x*,*j*) | –2.9287 | –0.6865 | 0.0421 | 0.3037 | 0.4236 | 0.4222 | 0.3486 |
| | *c*(*x*,*j*) | 0.8507 | –0.2745 | –0.5156 | –0.5656 | –0.5898 | –0.5456 | –0.4624 |
| Princeton 'South' | *a*(*x*,*j*) | 1.0819 | 1.2846 | 1.2223 | 1.1905 | 1.1911 | 1.1564 | 1.1307 |
| | *b*(*x*,*j*) | –3.0005 | –0.6181 | 0.0851 | 0.2631 | 0.3152 | 0.3017 | 0.2596 |
| | *c*(*x*,*j*) | 0.8689 | –0.3024 | –0.4704 | –0.4487 | –0.4291 | –0.3958 | –0.3538 |
| Princeton 'East' | *a*(*x*,*j*) | 1.1461 | 1.2231 | 1.1593 | 1.1404 | 1.1540 | 1.1336 | 1.1201 |
| | *b*(*x*,*j*) | –2.2536 | –0.4301 | 0.0581 | 0.1991 | 0.2511 | 0.2556 | 0.2362 |
| | *c*(*x*,*j*) | 0.6259 | –0.2245 | –0.3479 | –0.3487 | –0.3506 | –0.3428 | –0.3268 |
| Princeton 'West' | *a*(*x*,*j*) | 1.1415 | 1.2563 | 1.1851 | 1.1720 | 1.1865 | 1.1746 | 1.1639 |
| | *b*(*x*,*j*) | –2.7070 | –0.5381 | 0.0633 | 0.2341 | 0.3080 | 0.3314 | 0.3190 |
| | *c*(*x*,*j*) | 0.7663 | –0.2637 | –0.4177 | –0.4272 | –0.4452 | –0.4537 | –0.4435 |
| United Nations 'Latin America' | *a*(*x*,*j*) | 0.6892 | 1.3625 | 1.0877 | 0.7500 | 0.5605 | 0.5024 | 0.5326 |
| | *b*(*x*,*j*) | –1.6937 | –0.3778 | 0.0197 | 0.0532 | 0.0222 | 0.0028 | 0.0052 |
| | *c*(*x*,*j*) | 0.6464 | –0.2892 | –0.2986 | –0.1106 | 0.0170 | 0.0048 | 0.0256 |
| | *d*(*x*,*j*) | 0.0106 | –0.0041 | 0.0024 | 0.0115 | 0.0171 | 0.0180 | 0.0168 |
| United Nations 'Chilean' | *a*(*x*,*j*) | 0.8274 | 1.3129 | 1.0632 | 0.8236 | 0.6895 | 0.6098 | 0.5615 |
| | *b*(*x*,*j*) | –1.5854 | –0.2457 | 0.0196 | 0.0293 | 0.0068 | –0.0014 | 0.0040 |
| | *c*(*x*,*j*) | 0.5949 | –0.2329 | –0.1996 | –0.0684 | 0.0032 | 0.0166 | 0.0073 |
| | *d*(*x*,*j*) | 0.0097 | –0.0031 | 0.0021 | 0.0081 | 0.0119 | 0.0141 | 0.0159 |
| United Nations 'South Asian' | *a*(*x*,*j*) | 0.6749 | 1.3716 | 1.0899 | 0.7694 | 0.6156 | 0.6077 | 0.6952 |
| | *b*(*x*,*j*) | –1.7580 | –0.3652 | 0.0299 | 0.0548 | 0.0231 | 0.0040 | 0.0018 |
| | *c*(*x*,*j*) | 0.6805 | –0.2966 | –0.2887 | –0.0934 | 0.0298 | 0.0573 | 0.0306 |
| | *d*(*x*,*j*) | 0.0109 | –0.0041 | 0.0024 | 0.0108 | 0.0149 | 0.0141 | 0.0109 |
| United Nations 'Far Eastern' | *a*(*x*,*j*) | 0.7194 | 1.2671 | 1.0668 | 0.7833 | 0.5765 | 0.4115 | 0.3071 |
| | *b*(*x*,*j*) | –1.3143 | –0.2996 | 0.0017 | 0.0307 | 0.0068 | 0.0014 | 0.0111 |
| | *c*(*x*,*j*) | 0.5432 | –0.2105 | –0.2424 | –0.1103 | –0.0202 | 0.0083 | 0.0129 |
| | *d*(*x*,*j*) | 0.0093 | –0.0029 | 0.0019 | 0.0098 | 0.0165 | 0.0213 | 0.0251 |
| United Nations 'General' | *a*(*x*,*j*) | 0.7210 | 1.3115 | 1.0768 | 0.7682 | 0.5769 | 0.4845 | 0.4760 |
| | *b*(*x*,*j*) | –1.4686 | –0.3360 | 0.0109 | 0.0439 | 0.0176 | 0.0034 | 0.0071 |
| | *c*(*x*,*j*) | 0.5746 | –0.2475 | –0.2695 | –0.1090 | 0.0038 | 0.0036 | 0.0246 |
| | *d*(*x*,*j*) | 0.0095 | –0.0034 | 0.0021 | 0.0105 | 0.0165 | 0.0187 | 0.0189 |

Sources: Princeton models: UN Population Division (1983); UN Models: UN Population Division (1991)

***Step 5c: Estimate the time reference t(x) of each estimated ${}_nq_0$***

Using the model life table family $j$ identified as being appropriate, parameters $e(x,j)$, $f(x,j)$ and $g(x,j)$ are substituted into the following equation:

$$t(x) = e(x,j) + f(x,j) \times \frac{{}_5P_{15}}{{}_5P_{20}} + g(x,j) \times \frac{{}_5P_{20}}{{}_5P_{25}}.$$

The results are presented in Table 16.6.

*TABLE 16.2 Values of e(x,j), f(x,j) and g(x,j) for estimating the reference period t(x) for probabilities of dying by exact ages of childhood from proportions dead of children ever born classified by age of mother*

| Family $j$ | | Age group of mother and value of $n$ in ${}_nq_0$ | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 15–19 | 20–24 | 25–29 | 30–34 | 35–39 | 40–44 | 45–49 |
| | | 1 | 2 | 3 | 5 | 10 | 15 | 20 |
| Princeton 'North' | $e(x,j)$ | 1.0921 | 1.3207 | 1.5996 | 2.0779 | 2.7705 | 4.1520 | 6.9650 |
| | $f(x,j)$ | 5.4732 | 5.3751 | 2.6268 | –1.7908 | –7.3403 | –12.2448 | –13.9160 |
| | $g(x,j)$ | –1.9672 | 0.2133 | 4.3701 | 9.4126 | 14.9352 | 19.2349 | 19.9542 |
| Princeton 'South' | $e(x,j)$ | 1.0900 | 1.3079 | 1.5173 | 1.9399 | 2.6157 | 4.0794 | 7.1796 |
| | $f(x,j)$ | 5.4443 | 5.5568 | 2.6755 | –2.2739 | –8.4819 | –13.8308 | –15.3880 |
| | $g(x,j)$ | –1.9721 | 0.2021 | 4.7471 | 10.3876 | 16.5153 | 21.1866 | 21.7892 |
| Princeton 'East' | $e(x,j)$ | 1.0959 | 1.2921 | 1.5021 | 1.9347 | 2.6197 | 4.1317 | 7.3657 |
| | $f(x,j)$ | 5.5864 | 5.5897 | 2.4692 | –2.6419 | –8.9693 | –14.3550 | –15.8083 |
| | $g(x,j)$ | –1.9949 | 0.3631 | 5.0927 | 10.8533 | 17.0981 | 21.8247 | 22.3005 |
| Princeton 'West' | $e(x,j)$ | 1.0970 | 1.3062 | 1.5305 | 1.9991 | 2.7632 | 4.3468 | 7.5242 |
| | $f(x,j)$ | 5.5628 | 5.5677 | 2.5528 | –2.4261 | –8.4065 | –13.2436 | –14.2013 |
| | $g(x,j)$ | –1.9956 | 0.2962 | 4.8962 | 10.4282 | 16.1787 | 20.1990 | 20.0162 |
| United Nations 'Latin America' | $e(x,j)$ | 1.1703 | 1.6955 | 1.8296 | 2.1783 | 2.8836 | 4.4580 | 6.9351 |
| | $f(x,j)$ | 0.5129 | 4.1320 | 2.9020 | –2.5688 | –10.3282 | –17.1809 | –19.3871 |
| | $g(x,j)$ | –0.3850 | –0.1635 | 3.4707 | 9.0883 | 15.4301 | 20.4296 | 23.4007 |
| United Nations 'Chilean' | $e(x,j)$ | 1.3092 | 1.6897 | 1.8368 | 2.2036 | 2.9955 | 4.7734 | 7.4495 |
| | $f(x,j)$ | 1.9474 | 4.6176 | 2.6370 | –3.3520 | –11.4013 | –17.8850 | –19.0513 |
| | $g(x,j)$ | –0.7982 | –0.0173 | 4.0305 | 9.9233 | 16.3441 | 20.8883 | 23.0529 |
| United Nations 'South Asian' | $e(x,j)$ | 1.1922 | 1.7173 | 1.8631 | 2.1808 | 2.7654 | 4.1378 | 6.4885 |
| | $f(x,j)$ | 0.7940 | 4.3117 | 2.8767 | –2.7219 | –10.8808 | –18.6219 | –22.2001 |
| | $g(x,j)$ | –0.5425 | –0.1653 | 3.5848 | 9.3705 | 16.2255 | 22.2390 | 26.4911 |
| United Nations 'Far Eastern' | $e(x,j)$ | 1.2779 | 1.7471 | 1.9107 | 2.3172 | 3.2087 | 5.1141 | 7.6383 |
| | $f(x,j)$ | 1.5714 | 4.2638 | 2.7285 | –2.6259 | –9.8891 | –15.3263 | –15.5739 |
| | $g(x,j)$ | –0.6994 | –0.0752 | 3.5881 | 9.0238 | 14.7339 | 18.2507 | 19.7669 |
| United Nations 'General' | $e(x,j)$ | 1.2136 | 1.7025 | 1.8360 | 2.1882 | 2.9682 | 4.6526 | 7.1425 |
| | $f(x,j)$ | 0.9740 | 4.1569 | 2.8632 | –2.6521 | –10.3053 | –16.6920 | –18.3021 |
| | $g(x,j)$ | –0.5247 | –0.1232 | 3.5220 | 9.1961 | 15.3161 | 19.8534 | 22.4168 |

Sources: Princeton models: UN Population Division (1983); UN Models: UN Population Division (1991)

| Family *j* | | Time since first birth of mother and value of *n* in $_nq_0$ | | | | |
|---|---|---|---|---|---|---|
| | | 0–4 | 5–9 | 10–14 | 15–19 | 20–24 |
| | | 2 | 5 | 5 | 5 | 10 |
| Princeton 'North' | *a*(*x*,*j*) | 1.1980 | 1.2248 | 1.2076 | 1.2030 | 1.3292 |
| | *b*(*x*,*j*) | –0.1266 | –0.1919 | –0.0105 | 0.0896 | 0.1598 |
| | *c*(*x*,*j*) | 0.0038 | –0.0870 | –0.2911 | –0.4265 | –0.5778 |
| Princeton 'South' | *a*(*x*,*j*) | 1.1705 | 1.3166 | 1.2952 | 1.2836 | 1.5269 |
| | *b*(*x*,*j*) | –0.1461 | –0.3157 | –0.0423 | 0.1308 | 0.2659 |
| | *c*(*x*,*j*) | 0.0051 | –0.0971 | –0.4295 | –0.6496 | –0.9174 |
| Princeton 'East' | *a*(*x*,*j*) | 1.2182 | 1.2769 | 1.2731 | 1.2585 | 1.3410 |
| | *b*(*x*,*j*) | –0.1809 | –0.2268 | 0.0005 | 0.1216 | 0.1749 |
| | *c*(*x*,*j*) | 0.0214 | –0.1052 | –0.3720 | –0.5013 | –0.5964 |
| Princeton 'West' | *a*(*x*,*j*) | 1.2049 | 1.2573 | 1.2431 | 1.2469 | 1.4258 |
| | *b*(*x*,*j*) | –0.1553 | –0.2266 | –0.0230 | 0.0999 | 0.1948 |
| | *c*(*x*,*j*) | 0.0135 | –0.0944 | –0.3409 | –0.5267 | –0.7454 |

*TABLE 16.3 Values of a(x,j), b(x,j), c(x,j) for estimating probabilities of dying by exact ages of childhood from proportions dead of children ever born classified by time since first birth of mother*

| Family *j* | | Time since first birth of mother and value of *n* in $_nq_0$ | | | | |
|---|---|---|---|---|---|---|
| | | 0–4 | 5–9 | 10–14 | 15–19 | 20–24 |
| | | 2 | 5 | 5 | 5 | 10 |
| Princeton 'North' | *e*(*x*,*j*) | 1.71 | 2.16 | 0.66 | –1.96 | –3.85 |
| | *f*(*x*,*j*) | 1.07 | 4.36 | 3.50 | –0.90 | –6.42 |
| | *g*(*x*,*j*) | –0.35 | 0.12 | 6.65 | 17.66 | 28.94 |
| Princeton 'South' | *e*(*x*,*j*) | 1.68 | 2.29 | 1.19 | –1.01 | –2.68 |
| | *f*(*x*,*j*) | 0.96 | 3.84 | 3.45 | –0.18 | –5.06 |
| | *g*(*x*,*j*) | –0.32 | –0.01 | 5.41 | 15.03 | 25.21 |
| Princeton 'East' | *e*(*x*,*j*) | 1.68 | 2.19 | 0.71 | –1.96 | –4.06 |
| | *f*(*x*,*j*) | 0.99 | 4.28 | 3.63 | –0.71 | –6.35 |
| | *g*(*x*,*j*) | –0.33 | 0.02 | 6.36 | 17.42 | 29.14 |
| Princeton 'West' | *e*(*x*,*j*) | 1.70 | 2.20 | 0.86 | –1.46 | –2.97 |
| | *f*(*x*,*j*) | 1.03 | 4.20 | 3.47 | –0.69 | –5.80 |
| | *g*(*x*,*j*) | –0.34 | 0.06 | 6.21 | 16.49 | 26.65 |

*TABLE 16.4 Values of e(x,j), f(x,j) and g(x,j) for estimating the reference period t(x) for probabilities of dying by exact ages of childhood from proportions dead of children ever born classified by time since first birth of mother*

*TABLE 16.5 (below) Values of a(x,j), b(x,j), c(x,j) for estimating probabilities of dying by exact ages of childhood from proportions dead of children ever born classified by duration of marriage of mother*

| Family *j* | | Duration of marriage of mother and value of *n* in $_nq_0$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0–4 | 5–9 | 10–14 | 15–19 | 20–24 | 25–29 |
| | | 2 | 3 | 5 | 10 | 15 | 20 |
| Princeton 'North' | *a*(*x*,*j*) | 1.2615 | 1.1957 | 1.3067 | 1.4701 | 1.5039 | 1.4798 |
| | *b*(*x*,*j*) | –0.5340 | –0.4103 | –0.0103 | 0.1763 | 0.0039 | –0.2487 |
| | *c*(*x*,*j*) | 0.1252 | –0.0930 | –0.4618 | –0.7268 | –0.7071 | –0.5582 |
| Princeton 'South' | *a*(*x*,*j*) | 1.3103 | 1.2309 | 1.2774 | 1.3493 | 1.3592 | 1.3532 |
| | *b*(*x*,*j*) | –0.5856 | –0.3463 | 0.0336 | 0.1366 | –0.0315 | –0.1978 |
| | *c*(*x*,*j*) | 0.1367 | –0.1073 | –0.3987 | –0.5403 | –0.4944 | –0.4099 |
| Princeton 'East' | *a*(*x*,*j*) | 1.2299 | 1.1611 | 1.2036 | 1.2773 | 1.3014 | 1.3160 |
| | *b*(*x*,*j*) | –0.3998 | –0.2451 | 0.0171 | 0.1015 | –0.0219 | –0.1630 |
| | *c*(*x*,*j*) | 0.0910 | –0.0797 | –0.2992 | –0.4276 | –0.4195 | –0.3751 |
| Princeton 'West' | *a*(*x*,*j*) | 1.2584 | 1.1841 | 1.2446 | 1.3353 | 1.3875 | 1.4227 |
| | *b*(*x*,*j*) | –0.4683 | –0.3006 | 0.0131 | 0.1157 | –0.0193 | –0.1954 |
| | *c*(*x*,*j*) | 0.1080 | –0.0892 | –0.3555 | –0.5245 | –0.5472 | –0.5127 |

***Step 6: Convert each estimate of ${}_nq_0$ into an estimate of ${}_5q_0$***

In the applications of the indirect child mortality estimation method presented here, each of the probabilities of dying by exact ages of childhood, ${}_nq_0$, is converted into a value of $\alpha$, the level parameter of a system of relational logit model life tables. The $\alpha$ is then used to estimate the corresponding probability of dying between birth and exact age 5, ${}_5q_0$:

$$\alpha = 0.5\left(\ln\left(\frac{{}_nq_0}{1-{}_nq_0}\right)\right) - Y^s(n)$$

where the estimates of ${}_nq_0$ come from Step 4 and the $Y^s(n)$ values are logit transformations of the standard life table. Thus, one obtains a series of values of $\alpha$ corresponding to the probabilities of dying estimated from data on the different age groups of respondents. Then for each $\alpha$:

$${}_5\hat{q}_0 = \frac{e^{2(\alpha+Y^s(5))}}{1+e^{2(\alpha+Y^s(5))}}.$$

To apply the relational model approach, it is necessary to choose a standard life table. In order to apply the indirect estimation procedure, it is necessary for Steps 4 and 5 to identify an appropriate model pattern, and the standard should be drawn from the same model family. The precise level of mortality within the family is less important than the family itself (the appropriate selection of which allows $\beta$ to be assumed to be 1 in the relational logit model life table). We therefore recommend choosing a standard with a life expectancy of 60 years.

***Step 7: Identify and interpret the results***

The resulting estimates of ${}_nq_0$ for each age group, the corresponding estimates of ${}_5q_0$, the time location estimates, and the time trend in ${}_5q_0$ must then be assessed. Plotted against time, the series of values of ${}_5q_0$ will give an indication of the time trend in levels of child mortality. If data from more than one census or survey are available, estimates of ${}_5q_0$ can be compared for the same time periods to evaluate the consistency and reliability of the data.

## WORKED EXAMPLE

The example uses data on children ever born and children surviving by age of mother from the 2008 Census of Malawi. The method is implemented in the associated Excel workbook (see website).

***Step 1: Calculate proportions dead of children ever born, ${}_5PD_x$***

Table 16.7 shows the basic data on number of women, number of children ever born, and number of children

*TABLE 16.6 Values of e(x,j), f(x,j) and g(x,j) for estimating reference period t(x) for probabilities of dying by exact ages of childhood from proportions dead of children ever born classified by duration of marriage of mother*

| Family *j* | | Duration of marriage of mother and value of *n* in ${}_nq_0$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0–4 | 5–9 | 10–14 | 15–19 | 20–24 | 25–29 |
| | | 2 | 3 | 5 | 10 | 15 | 20 |
| Princeton 'North' | *e(x,j)* | 1.03 | 1.70 | 1.43 | –0.08 | –1.97 | –2.19 |
| | *f(x,j)* | 1.31 | 4.21 | 3.27 | –1.08 | –3.48 | 0.61 |
| | *g(x,j)* | –0.33 | –0.02 | 4.41 | 12.93 | 21.33 | 23.94 |
| Princeton 'South' | *e(x,j)* | 1.02 | 1.66 | 1.21 | –0.65 | –2.91 | –3.16 |
| | *f(x,j)* | 1.31 | 4.51 | 3.47 | –1.60 | –4.14 | 1.21 |
| | *g(x,j)* | –0.33 | –0.03 | 4.95 | 14.68 | 24.01 | 26.35 |
| Princeton 'East' | *e(x,j)* | 1.04 | 1.64 | 1.11 | –0.86 | –3.22 | –3.39 |
| | *f(x,j)* | 1.42 | 4.70 | 3.30 | –1.97 | –4.11 | 1.67 |
| | *g(x,j)* | –0.35 | 0.06 | 5.45 | 15.52 | 24.86 | 26.98 |
| Princeton 'West' | *e(x,j)* | 1.03 | 1.67 | 1.21 | –0.54 | –2.47 | –2.21 |
| | *f(x,j)* | 1.37 | 4.59 | 3.33 | –1.77 | –3.92 | 1.31 |
| | *g(x,j)* | –0.34 | 0.02 | 5.14 | 14.64 | 23.10 | 24.45 |

surviving by five-year age group of mother from the 2008 Malawi Census. The proportion dead of children ever born ${}_5PD_x$ is calculated by dividing the number of surviving children by the number of children ever born, and subtracting the result from 1:

$$ {}_5PD_x = 1 - \frac{{}_5CS_x}{{}_5CEB_x}. $$

The results are shown in the fifth column of Table 16.7.

***Step 2: Calculate average numbers of children ever born to women in each five-year group, ${}_5P_x$***

Although the average number of children born to women in each five-year age group, ${}_5P_x$, is only needed for the age groups 15–19, 20–24 and 25–29, it is recommended to calculate these values for all age groups as part of data evaluation. The average is calculated simply by dividing children ever born by the number of women in the age group:

$$ {}_5P_x = \frac{{}_5CEB_x}{{}_5N_x}. $$

The results are shown in the sixth column of Table 16.7. The required parity ratios are then calculated as follows:

$$ \frac{{}_5P_{15}}{{}_5p_{20}} = \frac{0.2833}{1.5316} = 0.1850 $$

and

$$ \frac{{}_5P_{20}}{{}_5P_{25}} = \frac{1.5316}{2.8487} = 0.5376. $$

***Step 3: Select a model life table family***

A single set of proportions dead by age of mother contains essentially no information about the age pattern of child mortality in a population. However, in an actual country analysis, there would invariably be some other relevant information to guide a choice. The ideal information is the age pattern of child mortality from a full birth history for the same population. The comparison is of ${}_1q_0$ and ${}_4q_1$, and it is generally made graphically, superimposing observed points over curves showing the relationship at different mortality levels in each model family.

Several full birth history surveys have been conducted in Malawi. Figure 16.1 plots direct estimates for the 0 to 4 years before the DHS surveys of 1992, 2000 and 2004 (and also the 5 to 9 years before the 2000 survey) against model patterns. (Note that only three of the United Nations patterns are shown. This is because the General, South Asian, and Far Eastern patterns are indistinguishable in their age pattern of mortality under age 5.) All the observations show higher ${}_4q_1$ relative to ${}_1q_0$ than any of the model patterns. The optimal model choice in this instance would probably be the Princeton 'North' family, and it is this model that we use in what follows.

***Step 4a(1): Estimate the mean age of the age-specific fertility schedule***

We are using the Princeton 'North' family of model life tables, but for illustrative purposes, we estimate the mean age of the fertility schedule. The 2008 Malawi Census included a question for women of reproductive age on how many births they had in the year before the census. Table 16.8 shows the basic data, and the calculation of $\overline{m}$ . $\overline{m}$ is then calculated as the sum of column (vi) divided by the sum of column (iv), = 29.817/1.0375 = 28.74.

*TABLE 16.7 Children ever born and children surviving, Malawi, 2008 Census*

| Age group *x,x+4* | Number of women | Children ever born | Children surviving | Mean children ever born | Mean children surviving |
|---|---|---|---|---|---|
| 15–19 | 635,927 | 180,178 | 161,541 | 0.2833 | 0.2540 |
| 20–24 | 678,071 | 1,038,556 | 919,584 | 1.5316 | 1.3562 |
| 25–29 | 566,350 | 1,613,374 | 1,398,776 | 2.8487 | 2.4698 |
| 30–34 | 405,602 | 1,697,566 | 1,426,516 | 4.1853 | 3.5170 |
| 35–39 | 298,004 | 1,553,676 | 1,266,514 | 5.2136 | 4.2500 |
| 40–44 | 221,274 | 1,335,242 | 1,043,357 | 6.0343 | 4.7152 |
| 45–49 | 174,875 | 1,128,423 | 851,048 | 6.4527 | 4.8666 |

***Step 4a(2): Estimate ${}_nq_0$ from each ${}_5PD_x$***

Each ${}_5PD_x$ is then converted into an estimated ${}_nq_0$ using the appropriate coefficients from Table 16.4, as shown in Table 16.9. Thus for the age group 30–34,

$${}_5q_0 = 0.1597.\begin{pmatrix} 1.2046 + (0.3037 \times 0.1850) \\ + (-0.5656 \times 0.5376) \end{pmatrix} = 0.1528.$$

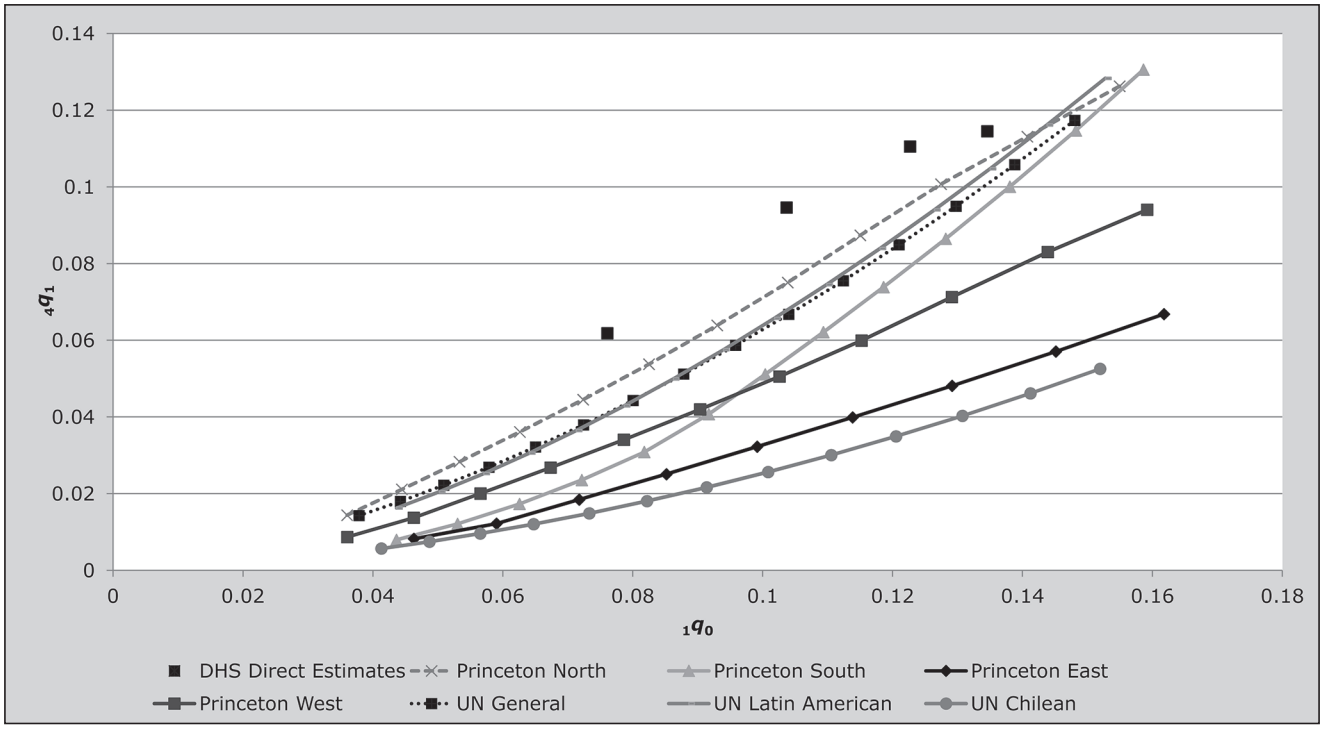
***Step 5: Estimate the time reference t(x) of each estimated ${}_nq_0$***

The time reference $t(x)$ of each estimate before the survey or census is then obtained, using the appropriate coefficients from Table 16.4, as shown in Table 16.10. Thus for the age group 30–34,

*TABLE 16.8 Calculation of $\overline{m}$ from births data from Malawi, 2008 Census*

| Age group $x,x+4$ at census | Number of women | Births in previous 12 months | Age-specific fertility rates, ${}_5f_x$ | Mid-point of age group $(x+2)$ | ${}_5f_x.(x+2)$ |
|---|---|---|---|---|---|
| 15–19 | 635,927 | 70,737 | 0.1112 | 17 | 1.891 |
| 20–24 | 678,071 | 169,406 | 0.2498 | 22 | 5.496 |
| 25–29 | 566,350 | 130,331 | 0.2301 | 27 | 6.213 |
| 30–34 | 405,602 | 79,232 | 0.1953 | 32 | 6.251 |
| 35–39 | 298,004 | 43,747 | 0.1468 | 37 | 5.432 |
| 40–44 | 221,274 | 13,956 | 0.0721 | 42 | 3.029 |
| 45–49 | 174,875 | 5,599 | 0.0320 | 47 | 1.505 |
| **Sum** | | | **1.0375** | | **29.817** |

*FIGURE 16.1 Direct estimates of ${}_4q_1$ and ${}_1q_0$ from Malawi DHSs, and the relationships in Princeton and UN model life tables*



$$t(30) = 2.0779 + (-1.7908) \times 0.1850 + 9.4126 \times 0.5377 = 6.81.$$

The Census was taken between 8 and 28 June 2008, so the estimated reference date can be found by subtracting $t$ from 2008.46 (the decimal year representation of 18 June 2008). Results are shown in the last column of Table 16.10.

***Step 6: Convert each estimate of ${}_nq_0$ into an estimate of ${}_5q_0$***

The final step is to convert each estimated ${}_nq_0$ into an estimate of the common index ${}_5\hat{q}_0$. This will make it possible to compare estimates across age groups. Each ${}_nq_0$ is converted into its logit $Y(n)$ by means of the identity $Y(n) = 0.5.ln({}_nq_0/(1-{}_nq_0))$. The value of $\alpha$ is then found by subtracting the standard logit $Y^s(n)$, from the North joint-sex model life tables with an expectation of life at birth of 60 years (column (vi)) from $Y(n)$. Each $\alpha$ is then used with the standard $Y^s(5)$ to get the estimated ${}_5\hat{q}_0$. Thus for the age group 25–29, $Y(3) = 0.5.\ln(0.1222/(1-0.1222)) = -0.9857$, and $\alpha = -0.9857 - (-1.1664) = 0.1806$. Then,

$${}_5\hat{q}_0 = \frac{e^{2(0.1806+(-1.0900))}}{1+e^{2(0.1806+(-1.0900))}} = 0.1396.$$

Estimates of ${}_1\hat{q}_0$ are derived in analogous fashion using the standard logit for age 1.

Figure 16.2 plots each estimate of ${}_1\hat{q}_0$ and ${}_5\hat{q}_0$ against the corresponding reference date. The figure indicates a declining trend in under-five mortality over time, from around 185 per 1000 in the early 1990s to around 140 in 2005. The apparent uptick in child mortality in 2007 is to be disregarded owing to the likely exaggeration of mortality estimated from very young mothers, as discussed above.

## DIAGNOSTICS, ANALYSIS AND INTERPRETATION

### Checks and validation

Regardless of how data have been collected, or of one's knowledge of how thoroughly interviewers were trained and supervised, careful data quality review is an essential first step of any analysis. All data sets contain errors, which can result from many sources, such as an interviewer cutting corners or an interviewee simply not knowing the correct answer to a question.

| Age group | Proportion dead of CEB | Regression coefficients for ${}_nq_0$ (Princeton 'North' Model) | | | ${}_nq_0$ |
|---|---|---|---|---|---|
| | | $a(i)$ | $b(i)$ | $c(i)$ | |
| 15–19 | 0.1034 | 1.1119 | –2.9287 | 0.8507 | 0.1063 |
| 20–24 | 0.1146 | 1.2390 | –0.6865 | –0.2745 | 0.1105 |
| 25–29 | 0.1330 | 1.1884 | 0.0421 | –0.5156 | 0.1222 |
| 30–34 | 0.1597 | 1.2046 | 0.3037 | –0.5656 | 0.1528 |
| 35–39 | 0.1848 | 1.2586 | 0.4236 | –0.5898 | 0.1885 |
| 40–44 | 0.2186 | 1.2240 | 0.4222 | –0.5456 | 0.2205 |
| 45–49 | 0.2458 | 1.1772 | 0.3486 | –0.4624 | 0.2441 |

*TABLE 16.9 Estimation of ${}_nq_0$ from each ${}_5PD_x$, Malawi, 2008 Census*

| | Regression coefficients for time ago: Princeton 'North' model | | | Time ago $t$ | Reference date |
|---|---|---|---|---|---|
| Age group | $e(i)$ | $f(i)$ | $g(i)$ | | (2008.46–$t$) |
| 15–19 | 1.0921 | 5.4732 | –1.9672 | 1.05 | 2007.42 |
| 20–24 | 1.3207 | 5.3751 | 0.2133 | 2.43 | 2006.03 |
| 25–29 | 1.5996 | 2.6268 | 4.3701 | 4.44 | 2004.03 |
| 30–34 | 2.0779 | –1.7908 | 9.4126 | 6.81 | 2001.66 |
| 35–39 | 2.7705 | –7.3403 | 14.9352 | 9.44 | 1999.02 |
| 40–44 | 4.1520 | –12.2448 | 19.2349 | 12.23 | 1996.24 |
| 45–49 | 6.9650 | –13.9160 | 19.9542 | 15.12 | 1993.35 |

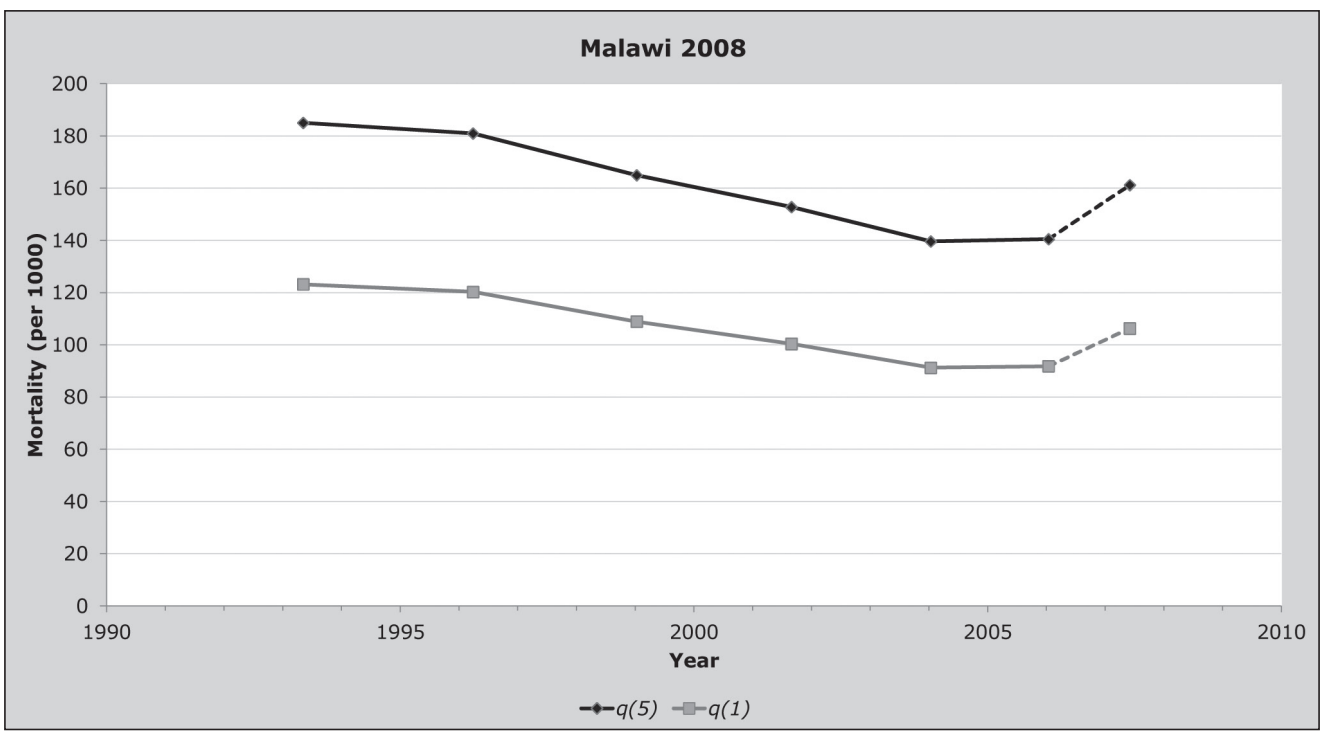
*TABLE 16.10 Estimation of time reference t(x) of each estimate in years before the census, Malawi, 2008 Census*

One specific aspect of summary birth history data needs to be emphasized for the data checking, and that is the failure of assumption 2, that mortality risks of children for a particular time period do not vary by age of mother. In many applications it is abundantly clear that this assumption does not hold. The risks for children of women aged 15–19 (and the indirect estimate of child mortality based on CEB and CD for this age group) are frequently higher, sometimes very substantially so, than the population average. The same is true to a lesser extent for the children of mothers aged 20–24. Two factors account for this distortion: the distribution of children by birth order and socio-economic factors. First births are known to be at higher risk of dying than higher-order births, and the children born to younger women clearly consist of an above-average proportion of first births. Women having children at early ages also tend to come from below-average socio-economic groups, and their children are thus exposed to above-average mortality

*TABLE 16.11 Estimating the logit life table parameter $\alpha$ for each estimate, and deriving a set of ${}_1\hat{q}_0$ and ${}_5\hat{q}_0$, Malawi, 2008 Census*

| Age group | ${}_nq_0$ | $n$ | logit $Y(n)$ | Standard logit $Y^s(n)$ | $\alpha$ | ${}_1\hat{q}_0$ | ${}_5\hat{q}_0$ |
|---|---|---|---|---|---|---|---|
| 15–19 | 0.1063 | 1 | –1.0647 | –1.3300 | 0.2653 | 0.1063 | 0.1612 |
| 20–24 | 0.1105 | 2 | –1.0431 | –1.2273 | 0.1842 | 0.0918 | 0.1405 |
| 25–29 | 0.1222 | 3 | –0.9857 | –1.1664 | 0.1806 | 0.0912 | 0.1396 |
| 30–34 | 0.1528 | 5 | –0.8566 | –1.0900 | 0.2334 | 0.1004 | 0.1528 |
| 35–39 | 0.1885 | 10 | –0.7299 | –1.0091 | 0.2791 | 0.1089 | 0.1650 |
| 40–44 | 0.2205 | 15 | –0.6313 | –0.9664 | 0.3350 | 0.1203 | 0.1809 |
| 45–49 | 0.2441 | 20 | –0.5652 | –0.9138 | 0.3487 | 0.1232 | 0.1850 |

*FIGURE 16.2 Estimated under-five and under-one mortality over time, Malawi, 2008 Census*



risks. Consequently the estimates of mortality derived from women aged 15–19 should be treated with a high degree of circumspection, or ignored altogether.

**Interpretation**

Two key characteristics of this method need to be borne in mind when interpreting results. First, there is no information about specific dates or ages in the basic data. The only timing information in the number of children ever borne by a woman is that those births occurred at some point between when she had her first birth and her age at the time of the survey. Even less information is available from the number of those children that have died about ages at death, since possible age ranges depend on time distributions of births. It is thus impossible to draw conclusions about short-term fluctuations in child mortality from a SBH. Reports of two women of the same age (duration of marriage, time since first birth) reporting the same numbers of children ever born and died can reflect experiences of different mortality conditions. The best that a summary birth history can offer is a broad indication of an average past trend. Even this average trend needs to be interpreted with care for the recent past because of the selection biases affecting reports of women aged 15–19 and, though to a lesser extent, 20–24.

The second characteristic is that information is provided only for surviving women who still live in surveyed households, with an associated risk of respondent selection bias. The mortality experience of children born in a community whose mothers no longer live in the community will not be included in the measures. If such children have higher mortality than those born to mothers who do still live in the community, mortality will be under-estimated. The most severe form of this bias is likely to result from substantial levels of HIV prevalence in the community, since such prevalence in the absence of widespread antiretroviral therapy will result in a strong positive correlation between survival of child and survival of mother (described in the section on the effects of HIV on methods for child mortality estimation in Chapter 15). However, some positive correlation between mother and child survival is almost certain in any population.

There are also other possible sources of respondent selection bias. For example, high in-migration rates will result in women reporting on the survival of children born and raised elsewhere, while high out-migration will remove responses about children who were born and raised in the community. Though it is impossible to know *a priori* the direction or magnitude of such biases, the analyst needs to keep in mind their potential effect. Non-response itself may actually be a smaller problem for the summary birth history than the full birth history, since information is often collected from third parties, not necessarily from the woman herself. Thus a summary birth history for a woman absent from the community on an extended trip might be collected when a full birth history for the same woman would not be since she could not be interviewed in person.

## DETAILED DESCRIPTION OF METHOD

The idea that proportions of children dead among children ever born were indicators of child mortality has a long history. Questions on children ever born and children surviving were included in the 1900 Census of the United States (Preston and Haines 1991), the 1911 Census of Britain, and the 1940 Census of Brazil, among others. However, the first methodology for translating such proportions into standard life table indicators was proposed by Brass and Coale (1968).

To illustrate the basic idea, take a simple if implausible example of a population in which all women have exactly one child, born at exact age 25, that all the women survive from age 25 to age 30, and that there is no migration. In a survey, the proportion dead among children borne by women of exact age 30 will precisely measure the cohort probability of dying between birth and exact age 5, ${}_5q_0$. In another population, all women also have exactly one child, but at age 27; in this population, the proportion dead will precisely measure the cohort probability of dying by age 3, ${}_3q_0$.

These two examples illustrate a number of important points. First, age of the women is a proxy for the exposure to risk of their children. Other things being equal, the older the mother, the longer on average her children have been exposed to the risk of dying. Second, the interpretation of a proportion dead of children ever born in terms of a standard life table measure depends on the age of childbearing. Third, equivalence of a proportion dead to a life table measure requires that there be no selection effects by mortality or out-migration, nor contamination effects by in-migration. Fourth, the measures obtained are for cohorts (or averages across cohorts) rather than for time periods.

Of course in real populations children are born at a range of ages of mother and are exposed to age-specific mortality risks that may change over time. Estimation methods use model age patterns of fertility and child mortality to create

model proportions dead of children ever born that can then be related to underlying life table parameters. A common feature of data on survival of children classified by age group of mother is that proportions dead for women aged 15–19 and 20–24 are higher than for subsequent age groups, despite the fact that they reflect shorter average exposure times of children. The problem arises because young women who have children are generally of below average socio-economic status, and their children are disproportionately first births, both of which are known risk factors for high child mortality. The experience of the children of such young mothers is therefore not representative of all children born in the population. Partly to address this bias, methods have been developed classifying women by duration of marriage (Sullivan 1972) and time since first birth (Hill and Figueroa 2001). It is also argued that these methods are less affected by fertility change.

The method was originally developed by Brass without explicit consideration of the effects on the estimates of mortality change, though he notes that under conditions of change "the estimates of $q(2)$ and $q(3)$ would be representative of the average mortality for a short period (less than a decade) before the census or survey" (Brass and Coale, 1968: 116). It is clear today that child mortality has been declining globally, and very rapidly in some populations. Following pioneering work by Feeney (1976, 1980), methods were developed to estimate a 'time reference' for the estimate derived from each age/duration group (Coale and Trussell (1977); Palloni and Heligman (1985); Hill and Figueroa (2001)). The proportion dead of CEB for a group of women represents an average of mortality risks across all the birth cohorts of their children. The older the women, or the longer their exposure, the further back in time the cohorts stretch, and the further back the reference time for a child mortality estimate based on the proportion dead. Since child mortality risks are highly concentrated at young ages, the reference date is not in practice very different from the average number of years ago that the births occurred if mortality trends have been fairly stable over time. The exact calendar year reference dates for the child mortality estimates thus have to be treated as central points in time with a distribution of deaths around them. As a result, the child mortality estimates derived from SBHs and variants of the Brass indirect methods cannot be used to identify mortality changes or crises located at a particular point in time. The method provides a good description of general trends in child mortality but smoothed in comparison with the true year-to-year fluctuations seen in almost every population. Other methods of measurement (such as FBH) are thus needed to estimate the temporal impact of, for example, health interventions on child survival.

## MATHEMATICAL EXPOSITION (*MANUAL X* AND UN MODEL LIFE TABLE METHODS)

The proportion dead of children ever borne by women of exact age $x$, $PD(x)$, is a birth-weighted average of cohort probabilities of dying,

$$PD(x) = \frac{\int_{\alpha}^{x} f(y)\,{}_{x-y}q_0{}^{c}\,dy}{\int_{\alpha}^{x} f(y)\,dy},$$

where $f(y)$ is the fertility rate at age $y$, ${}_{x-y}q^{c}{}_0$ is the probability of dying by age $(x-y)$ for the cohort born $(x-y)$ years earlier, and $\alpha$ is the earliest age of childbearing. The expression is exactly the same for duration of marriage and time since first birth, except that $\alpha$ becomes 0. Proportions dead for five-year age or duration groups can then be estimated by averaging point $PD$ estimates across the group, weighting for assumed population distributions $N_x$ at each $x$. It is typically assumed that the underlying population can be regarded as stable, with a growth rate appropriate for the demographic parameters underlying the $PD(x)$, and that the underlying life table at adult ages is that used in the calculation of the $PD(x)$. Such calculations use discrete age groups, for example, using single years of age,

$${}_5PD_x = \frac{\sum_{a=x}^{x+5} PD(a).{}_1L_a.e^{-r(a+0.5)}}{\sum_{a=x}^{x+5} {}_1L_a.e^{-r(a+0.5)}}.$$

Model schedules of $f$, $q$ and $N$ are used to model the ${}_5PD_x$ values, which are then related to appropriate $q$ values by regression analysis using parity ratios as the independent variables (see estimating equations above).

## EXTENSIONS OF THE METHOD

### Variants by duration of marriage or time since first birth

As described above, though not illustrated in detail, variants of the original method have been developed, classifying women by duration of marriage (Sullivan 1972) or time

since first birth (Hill and Figueroa 2001). The methods were developed to address two potential sources of error in the age-based method: effects of changing fertility (primarily through distorting the parity ratios) and the above-average mortality risks of children born to young mothers.

While these variants are illustrated below, the two refinements suffer from limitations of their own. First, many population censuses do not routinely collect the information needed to tabulate women and their children by duration of marriage or time since first birth. Second, with regard to the duration of marriage approach, in many developing countries marriage is not a prerequisite for commencement of childbearing. Further, the mortality of children of unmarried mothers may well be higher than that of mothers who are married. Thus distortions may also arise in the results of these variants. Where such data are collected, as in many Arab countries with low proportions of births outside marriage, they can provide important insights into the relative performance of the different approaches.

Some surveys, for instance the MICS (Multiple Indicator Cluster Surveys, conducted under the auspices of UNICEF) carried out before about 2010, have collected summary birth histories and the necessary information on date of marriage and first birth. The question thus arises as to which variant of the methodology should be preferred.

To explore this question, we have used a full birth history survey, the 1999–2000 Bangladesh Demographic and Health Survey. The data have been tabulated in all three formats, and analysed using the Princeton 'South' family of model life tables. Figure 16.3 shows the estimates of $_5q_0$ derived from each approach. (It may be noted that this particular data set does not show a strong over-estimation of mortality based on the most recent (15–19) age group.)

If one ignores in each series the most recent point (15–19 for age, 0–4 for both duration of marriage and time since first birth) the general trends of the estimates are very similar. The marital duration and time since first birth methods show similar levels, but from age group 25–29 onwards, the age estimates are always highest, averaging 20 per 1,000 higher (155 per 1,000) than either of the other two

***FIGURE 16.3*** *Estimates of under-five mortality derived from proportions dead of children ever born classified by age group, marital duration group and time since first birth group of mother, Bangladesh, 1999–2000 DHS*

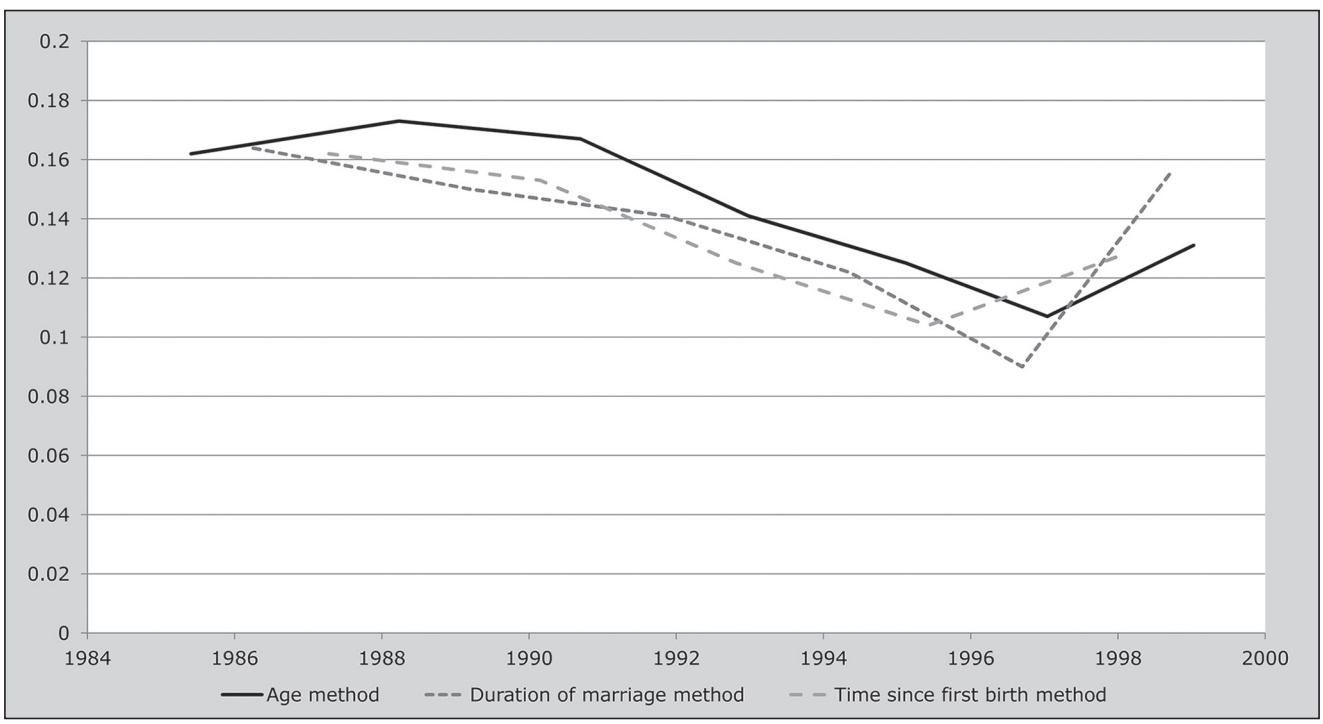
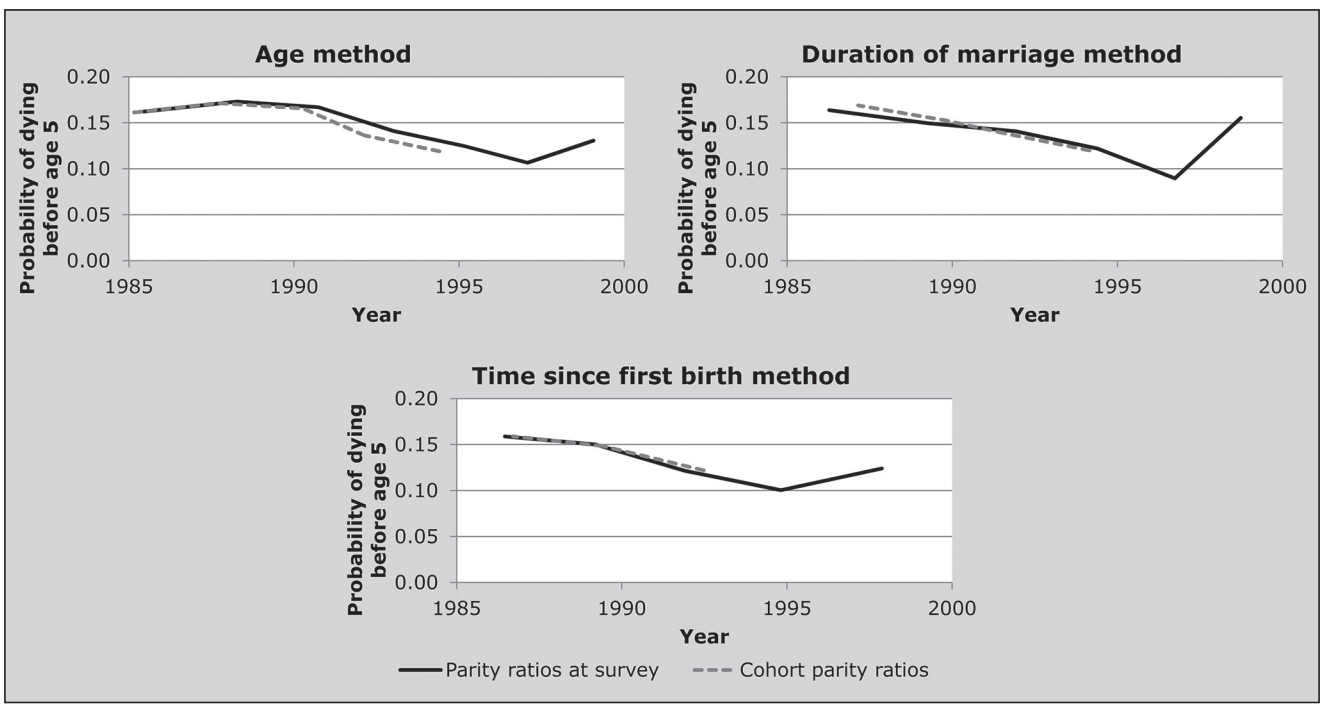


*TABLE 16.12 Parity ratios P(1)/P(2) and P(2)/P(3) calculated at the time of survey versus for true cohorts, Bangladesh, 1999–2000 DHS*

| Age method | | | | | | |
|---|---|---|---|---|---|---|
| Parity ratio | Time of survey | Cohort aged 25–29 | Cohort aged 30–34 | Cohort aged 35–39 | Cohort aged 40–44 | Cohort aged 45–49 |
| *P*(1)/*P*(2) | 0.269 | 0.342 | 0.311 | 0.327 | 0.361 | 0.336 |
| *P*(2)/*P*(3) | 0.544 | 0.660 | 0.639 | 0.603 | 0.632 | 0.619 |
| **Duration of marriage method** | | | | | | |
| Parity ratio | Time of survey | Cohort 10–14 | Cohort 15–19 | Cohort 20–24 | Cohort 25–29 | |
| *P*(1)/*P*(2) | 0.353 | 0.352 | 0.288 | 0.274 | 0.287 | |
| *P*(2)/*P*(3) | 0.635 | 0.682 | 0.650 | 0.598 | 0.603 | |
| **Time since first birth method** | | | | | | |
| Parity ratio | Time of survey | Cohort 10–14 | Cohort 15–19 | Cohort 20–24 | | |
| *P*(1)/*P*(2) | 0.553 | 0.544 | 0.495 | 0.505 | | |
| *P*(2)/*P*(3) | 0.709 | 0.775 | 0.753 | 0.723 | | |

*FIGURE 16.4 Indirect estimates of under-five mortality using parity ratios observed at the survey versus ratios for true cohorts, Bangladesh, 1999–2000 DHS*



methods. Straight averages of all the duration of marriage and time since first birth estimates for all points except the most recent give almost the same result, 133 for marriage duration and 136 from time since first birth, although the estimates span slightly different periods.

There are several points of interest in this application. First, the marriage duration and time since first birth methods were developed in part to try to circumvent the problem of selection bias in the age-based points for women aged 15–19 and 20–24. In this application, however, these two methods show jumps for the most recent point as large as or larger than that for the age-based estimates. This could be the case if the dominant selection bias is for first births, all (time since first birth) or almost all (marriage duration) of which will be concentrated in the first category, whereas they will be more spread out across age groups.

A second issue is the use of parity ratios as observed at the time of the survey to reflect the time distribution of births in the past. As noted above, if fertility is declining (even if the relative age pattern of fertility is not changing), the parities for the younger women will be lower and those for the older women higher, reducing the parity ratios below the values for any true cohort. This will give the appearance of more recent childbearing (and thus lower average duration of child exposure to the risk of dying) than is really the case, and thus lead to over-estimates of child mortality.

Bangladesh experienced a substantial fertility decline from the mid-1980s to the time of the 1999–2000 DHS, so it is interesting to explore the size of this possible bias. Since the data come from a full birth history, we can calculate cohort parities at points 5, 10, 15 etc. years in the past by subtracting recent births from children ever born, and then calculate parity ratios for true cohorts. We can only use the standard methodology (using ratios $P(1)/P(2)$ and $P(2)/P(3)$) for cohorts that have reached the third age or duration group, thus age groups 25–29 and up or time since first birth/marriage duration categories of 10–14 and up.

The first panel of Table 16.12 compares parity ratios at the time of survey and cohort parity ratios. The expected pattern is quite clear for the age method, with the time of survey ratios being clearly below any of the cohort ratios, which are themselves rather stable across cohorts. The picture is less clear for the marital duration and time since first birth methods however: the $P(2)/P(3)$ ratios are generally larger than the time of survey ratios, whereas the $P(1)/P(2)$ ratios are uniformly lower.

To get a sense of the magnitude of this effect on the estimates, we can use cohort parity ratios for each proportion dead of children ever born (for age groups 25–29 and over, and duration groups 10–14 and over). The three panels of Figure 16.4 show the original and cohort-based estimates by reference date, for age groups, marital duration groups and time since first birth groups respectively. The use of cohort ratios reduces the estimates based on age groups 25–29 and 30–34, but has little effect on the estimates from the other two methods. One possible explanation would be that the fertility decline in Bangladesh over this period was primarily an increase in age at childbearing, with small effects by duration of marriage or time since first birth. The net effect of using cohort ratios is to improve consistency across methods, but at the rather high cost of losing the most recent estimates.

The bottom line is that the age-based method will tend to over-estimate child mortality if fertility is changing rapidly, whereas the effects on the other two approaches appear to be small. If the data are available, it is advisable to use one of the other two variants rather than the age-based method, but it may not be worth including an extra question in a census just to get this information.

## FURTHER READING AND REFERENCES

The indirect estimation of child mortality is discussed in all the classic manuals on indirect estimation (Sloggett, Brass, Eldridge *et al*. 1994; UN Population Division 1983).


Brass W and AJ Coale. 1968. "Methods of analysis and estimation," in Brass, W, AJ Coale, P Demeny, DF Heisel, *et al.* (eds). *The Demography of Tropical Africa*. Princeton NJ: Princeton University Press, pp. 88–139.

Brass W, AJ Coale, P Demeny, DF Heisel *et al.* (eds). 1968. *The Demography of Tropical Africa*. Princeton NJ: Princeton University Press.

Coale AJ and J Trussell. 1977. "Estimating the time to which Brass estimates apply; Annex I to Preston SH and Palloni A 'Fine-tuning Brass-type mortality estimates with data on ages of surviving children'", *Population Bulletin of the United Nations* **10**:87–89.

Collumbien M and A Sloggett. 2001. "Adjustment methods for bias in the indirect childhood mortality estimates," in Zaba, B and J Blacker (eds). *Brass Tacks: Essays in Medical Demography*. London: Athlone, pp. 20–42.

Feeney G. 1976. "Estimating infant mortality rates from child survivorship data by age of mother", *Asian and Pacific Census Newsletter* **3**(2):12–16. http://hdl.handle.net/10125/3556

Feeney G. 1980. "Estimating infant mortality trends from child survivorship data", *Population Studies* **34**(1):109–128. doi: 10.1080/00324728.1980.10412839

Hill K and M-E Figueroa. 2001. "Child mortality estimation by time since first birth," in Zaba, B and J Blacker (eds). *Brass Tacks: Essays in Medical Demography*. London: Athlone, pp. 9–19.

Palloni A and L Heligman. 1985. "Re-estimation of structural parameters to obtain estimates of mortality in developing countries", *Population Bulletin of the United Nations* **18**:10–33.

Preston SH and MR Haines. 1991. *Fatal Years: Child Mortality in Late Nineteenth-century America.* Princeton, NJ: Princeton University Press.

Sloggett A, W Brass, SM Eldridge, IM Timæus, P Ward and B Zaba (eds). 1994. *Estimation of Demographic Parameters from Census Data*. Tokyo: Statistical Institute for Asia and the Pacific.

Sullivan JM. 1972. "Models for the estimation of the probability of dying between birth and exact ages of early childhood", *Population Studies* **26**(1):79–97. doi: 10.1080/00324728.1972.10405204

UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

UN Population Division. 1991. *Child Mortality in Developing Countries: Socio-economic Differentials, Trends and Implications*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/123.

# Survey Data Methods

# Chapter 17 Direct estimation of child mortality from birth histories

*Kenneth Hill*

In this chapter, we focus on the use of data from full birth histories (FBH) or truncated birth histories (TBH) to estimate child mortality. The key characteristics of such data are that for each birth included, the date of birth, survival status and (if dead) age at death are recorded. Analysis of the data typically uses life table approaches. Indirect estimation of child mortality, and estimation of child mortality from survival of a recent birth, are covered in Chapters 16 and 18 respectively.

## DATA REQUIREMENTS AND ASSUMPTIONS

### Data required

For each woman of reproductive age (in some settings for cultural reasons information collection is limited to ever-married women):

- the name of each child born alive;
- the month and year of birth of each child;
- the child's sex (optional);
- whether the child is still alive; and
- if the child has died, the age at death (age at death in a DHS programme is collected in days for deaths that occur in the first 28 days of life, in months for deaths at ages one to 23 months, and in years thereafter).

### Important assumptions

- Children still alive and children dead are reported with similar accuracy.
- Dates of birth and ages at death are reported with reasonable accuracy.
- No correlation exists between mortality risks of children and survival rates of mothers (whether as a result of mortality or migration) in the population.

## CAVEATS AND WARNINGS

The dangers associated with working with directly-collected data arise from two sources. The first is the risk of survivor bias as only living mothers are asked the detailed birth histories used to generate the data. In situations where it is anticipated that deceased mothers might have had different fertility, or different mortality among their children, from surviving mothers, there is a risk of appreciable bias in the estimates derived. Aspects of survivor bias are discussed in Chapter 15 in the section on the effects of HIV on child mortality estimation.

The second danger is that if an upper age limit is applied to the women from whom detailed birth history data are collected, truncation bias becomes more significant the further back in time one looks. If an age limit of 49 is applied to the collection of the data, this means that for the period 10 years before the survey, information is only available for women who were then aged up to age 39. Hence, child mortality estimated from such data for earlier time periods will be increasingly based on the experience of younger women. In turn, this might lead to measurement bias, as this truncation results in an over-representation of first births among younger women, meaning that child mortality thus estimated is likely to be increasingly overestimated for earlier time periods. There is some evidence that such over-estimation is counter-balanced by underestimation arising from recall bias (and selective omission of children who have died in periods longer in the past).

## DATA EVALUATION AND DATA ANALYSIS

Regardless of how data have been collected, or of one's knowledge of how thoroughly interviewers were trained and supervised, careful review of data quality is an essential first step of any analysis. All data sets contain errors. These can result from many sources, such as an interviewer cutting corners or an interviewee simply not knowing the correct answer to a question. Each section below starts with a description of data evaluation techniques before progressing to analysis

methods. These evaluation techniques examine both internal consistency within a data set, and external consistency with other data sets for the same population. It should be noted in passing that the presence of data errors does not necessarily mean that a data set should not be analysed; the important thing is to know how large the errors are, and take them into account when interpreting the findings.

**The full birth history: data quality assessment**

The first step in a thorough data quality assessment is to examine the extent of missing values. In an FBH, values may be missing for a number of reasons. For example, whole households included in the original sampling frame may be missing. Further, eligible women within interviewed households may have no data because the woman could not be interviewed. In addition, individual items within an FBH may be missing because the interviewed woman did not know a child's birth date, or whether a child was still alive, or (if the child had died) the age at death. The proportions of events potentially affected by these errors need to be examined. Missing items may be imputed during data cleaning, but imputed values should be flagged. The absence of missing values should not be taken as strong evidence of data quality, and may in fact be taken as a warning flag: in some surveys, interviewers and supervisors are trained to avoid missing values, and in such cases data may be more or less made up by the interviewer.

The second step in the data quality assessment is to examine the aggregate results for implausible irregularities. The irregularities most often identified are in sex ratios at birth, in annual distributions of live births, and in ages at death. In the absence of intervention, sex ratios in human populations are generally in the range of 100 to 106 males per 100 females. Sex ratios for birth cohorts outside this range are probably indicative of error. Sex ratios that increase for cohorts born a longer time before the survey are particularly clear indicators of an error, in this case under-reporting of female births that occurred in the distant past.

In the absence of major positive or negative events, births will normally be fairly smoothly distributed by calendar year in that while seasonality is common, this should not affect the annual numbers. Possible errors can be identified by calculating 'birth ratios', defined as $\frac{2B_t}{B_{t-1}+B_{t+1}}$, where $B_t$ is the number of births reported in a given year, $t$. An error commonly found in DHS data sets has come to be called "birth transference". DHS surveys collect a substantial amount of additional data about children born since some cut-off date, usually 1 January of the calendar year five years before the survey. It is often the case that births that occurred in that year are reported as occurring in the previous year, presumably to reduce work load. This results in a deficit of births in the year following the cut-off, and a surplus in the year immediately before the cut-off. Birth ratios will highlight this error, since the birth ratio for the year starting with the cut-off will be low, and that for the preceding year will be high. Very often, this birth transference is greater for children who have died than for those who are still alive, so it is good practice to calculate separate ratios for surviving and dead children.

Irregularities in reporting ages at death can similarly be identified by calculating ratios of deaths at some age $x$ to the average number of deaths at ages $(x-1)$ and $(x+1)$. In DHS data sets, there is generally an excess of deaths at age 7 days, to a lesser extent at age 14 days, and at age 12 months.

DHS conveniently publishes these data quality indicators at aggregate (national) level in survey reports (often in Appendix C). Analysts wishing to carry out sub-national analyses will need to calculate indicators themselves.

The data quality indicators described above measure internal plausibility. However, data can be internally plausible and still wrong. Data should also be evaluated by comparison with other surveys for the same population. Cohort comparisons are particularly powerful, for example comparing the average number of children ever borne by women aged 30–34 reported in one survey to the average number borne by women aged 35–39 reported in another survey five years later. Similar comparisons can be made of average numbers of children dead. Sequences of births by single calendar year for overlapping periods can also be compared, though one has to bear in mind that births in the past are increasingly truncated in birth histories limited to women aged 15–49 at the time of the survey.

## THE FULL BIRTH HISTORY: CALCULATION OF CHILD MORTALITY INDICATORS FOR BIRTH COHORTS

Widely used indicators of child mortality are expressed as probabilities. Thus the Infant Mortality Rate (IMR) is (approximately – as conventionally defined, the IMR is infant deaths in a year divided by births in the year, a value which closely approximates ${}_1q_0$) the probability of dying by

exact age 1, ${}_1q_0$, and the Under-five Mortality Rate (U5MR) is the probability of dying by age 5, ${}_5q_0$. Strictly speaking, probabilities are real cohort measures, even though most life tables calculate synthetic cohort measures for specified time periods from age-period mortality rates. Calculating cohort probabilities from FBH data is very straightforward. For example, the cohort IMR for births in the 12 to 23 months before the survey is simply the number of such births that died before the age of 1 divided by the number of births. Similarly, the cohort U5MR for births 5 to 9 years before the survey is the number of such births reported to have died before exact age 5 divided by the number of such births. Figure 17.1 shows the Lexis diagram representation of the age-cohort probability of dying by age 1 for the cohort born in July 2001 (in grey), and the age-period mortality of 5-month olds in calendar year 2002 (the black rectangle, relating to the example used later in this section).

Table 17.1 shows the relevant numbers and calculations for age-cohort probability of dying by age 1 for the cohort born in the 12–23 months before the survey and for the probability of dying by age 5 for the cohort born in the 5 to 9 years before the survey, using data from the 2004 Malawi DHS.

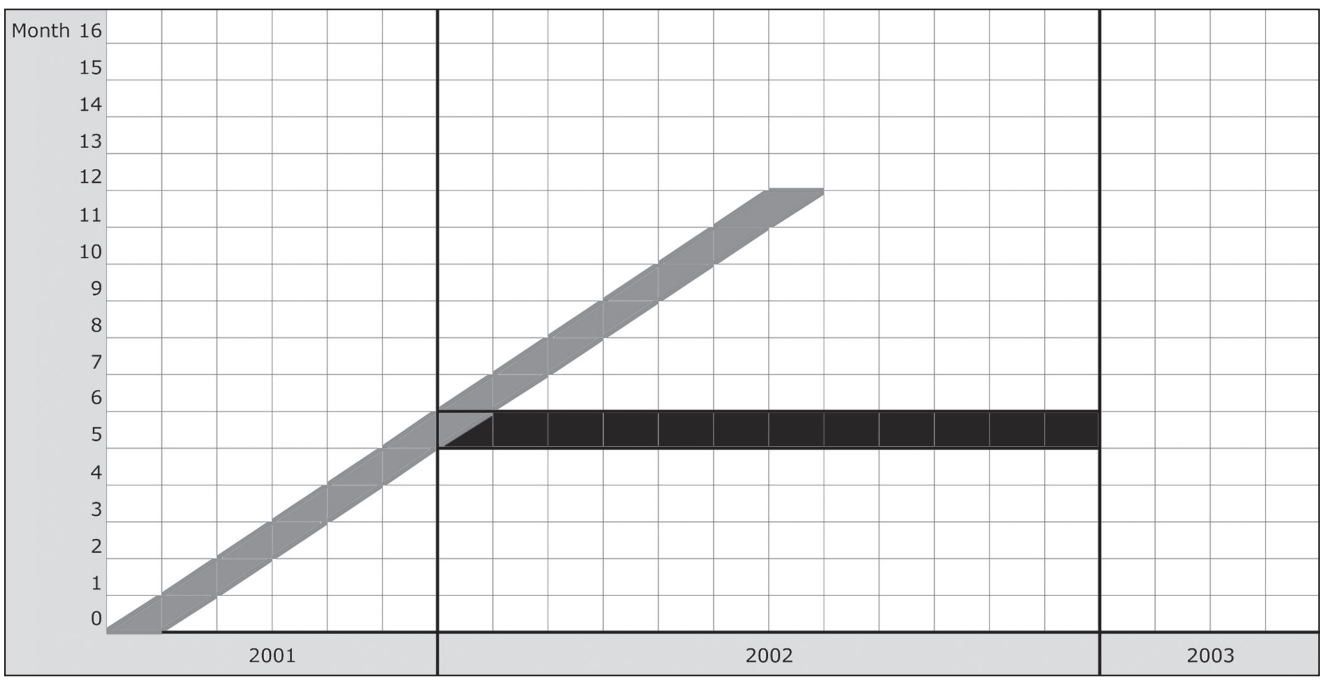
***TABLE 17.1*** *Calculation of IMR and U5MR for cohorts: Malawi, 2004 DHS*

| | Period of births | |
|---|---|---|
| | **12–23 months before the survey** | **60–119 months before the survey** |
| Births | 2,229 | 7,178 |
| Of which, deaths before 12 months | 143 | |
| Of which, deaths before 5 years | | 1,568 |
| Child mortality indicator | ${}_1q_0$ | ${}_5q_0$ |
| Cohort estimate per 1 000 births | 64.2 | 218.4 |

Note: weighted data; events in month of interview excluded

Note that there is no period interpretation of such cohort values; in the U5MR example, the cohort probability reflects mortality risks in every one of the 10 years before the survey. Also note that the probability of dying by age $x$ can only be calculated for cohorts that were born at least $x$ years before the survey. Both these considerations limit the value of the cohort measures, since for most purposes analysts and policy-makers are more interested in time period measures.

***FIGURE 17.1*** *Lexis diagram representation of age-cohort mortality by age 1 (in grey) and age-period mortality for 5-month olds (in black)*



## THE FULL BIRTH HISTORY: CALCULATION OF CHILD MORTALITY INDICATORS FOR TIME PERIODS

Period-specific measures are estimated using the synthetic cohort concept. Mortality rates for narrow age ranges and defined calendar periods are calculated on the basis of events and exposure in these rectangles in the Lexis Diagram. The rates are then converted into implied probabilities, using standard demographic relations (see, for example, Preston, Heuveline and Guillot (2001)) and making generally mild assumptions about the distribution of deaths in each rectangle. Finally, the probabilities of dying are applied successively to an initial hypothetical cohort of births to compute a survivorship curve $l(x)$ for each age $x$, from which it is easy to derive probabilities of dying.

FBH data lend themselves to these life table calculations quite easily. If data are collected following the standard DHS practice – as month and year of birth and age at death in days, months or years, depending on the age – deaths can be located with little ambiguity in age-period rectangles of the Lexis Diagram. (There will be some residual ambiguity, because of the imprecision of the information on date of birth and age at death, but the impact will depend on the sizes of the rectangles.) Here we describe an approach based on the calculation of age-specific mortality rates for a single calendar year (age-period rates) for mortality up to age 5. Extension to other time periods is straightforward. It is assumed that data are in standard DHS format, that is, birth dates are recorded in century month (CMC) format, and ages at death in days, months or years. Unit record data must be available. The unit of age used is the month. The basic calculations are therefore of age-specific mortality rates by month of age and calendar year. These rates are converted into corresponding probabilities of dying in each month. These probabilities are then converted into probabilities of surviving, and are chained together over whatever age range is required (typically up to age 5). The key to the calculation is to assign deaths and exposure time to one-month age segments across a calendar year.

### Data manipulation

Four variables in a DHS birth data set are required:

1) *b3*, date of birth in CMC;
2) *b5*, whether child is still alive;
3) *b6*, age at death, where the first digit represents the unit (1 indicating days; 2, months; and 3, years) and the second and third digits represent the value given that unit; and
4) *v005*, sample weight, expressed in millions.

Note that variable *b7*, age at death (months-imputed) is not used. This variable does not lend itself to the mortality rate approach described here because in cases in which age at death is recorded in years, the 'imputed' month is actually the lower bound of the age interval; that it, if age at death is recorded as '3 years', the imputed age at death in months is recorded as 36 months. Using this variable will result in systematic mis-location of deaths in time.

## APPLICATION OF METHOD

### *Step 1: Manipulation of age at death and calculation of estimated birth date and age at death*

We want to locate deaths in a calendar month of occurrence. Since we do not have a precise date of birth (only CMC), and in general we do not have a precise age at death (except for neonatal deaths), we need to impute both a date of birth and an age at death. We can perform this imputation using random numbers.

It is evidently undesirable – for reasons of lack of reproducibility, amongst others – to make use of a true random number generator to produce the random numbers referred to above. In addition, 'true' randomization risks creating a spurious impression of precision. As an alternative, we propose creating pseudo-random numbers from variables that are routinely available in DHS data and that can be applied in the algorithm above. It is an easy matter to create new variables apportioning the records into deciles based on the reported day of interview (*v016* in a DHS) and household number (*v002*). (These variables have been chosen on the grounds that there is unlikely to be any correlation between them and child mortality). These new variables will take the values in the range (0, 1 … 9). Dividing each by 10, and adding 0.05 results in two new uniformly distributed variables, *random1* and *random2*, taking values in the range (0.05, 0.15, … , 0.95).

It is then straightforward to impute a date of birth (*dob*, in months) if births in the month of interview are excluded from analysis by adding *random1* to *b3* (the CMC of the child's date of birth). The method for imputing an age at death (in units of months) depends on the 'unit'. For 'unit' = 1 (i.e. age at death measured in days), age at death (*aad*) can be estimated as ('value'+ *random2*)/31 (for age at death in days this is not necessary, but is described for symmetry);

for 'unit' = 2, age at death is 'value' + *random2*; and for 'unit' = 3, age at death is ('value' + *random2*)*12.

***Step 2: Location of deaths in target year***

For each month-of-age mortality rate, the events consist of deaths at that age in the period of investigation. Step 1 has imputed age at death in months. The date of death *dod* is given by the sum of imputed month of birth *dob* and imputed age of death *aad*. If imputed age at death is within the age range and the imputed date of death falls within the period of investigation, we have a relevant event.

***Step 3: Derivation of exposure to risk***

The calculation of exposure to risk is intricate, but relatively straightforward. The age range of the investigation refers to those ages (defined in appropriate units) for which we want to measure mortality. We define the lower bound of the age range to be $x_l$, and the upper bound to be $x_h$.

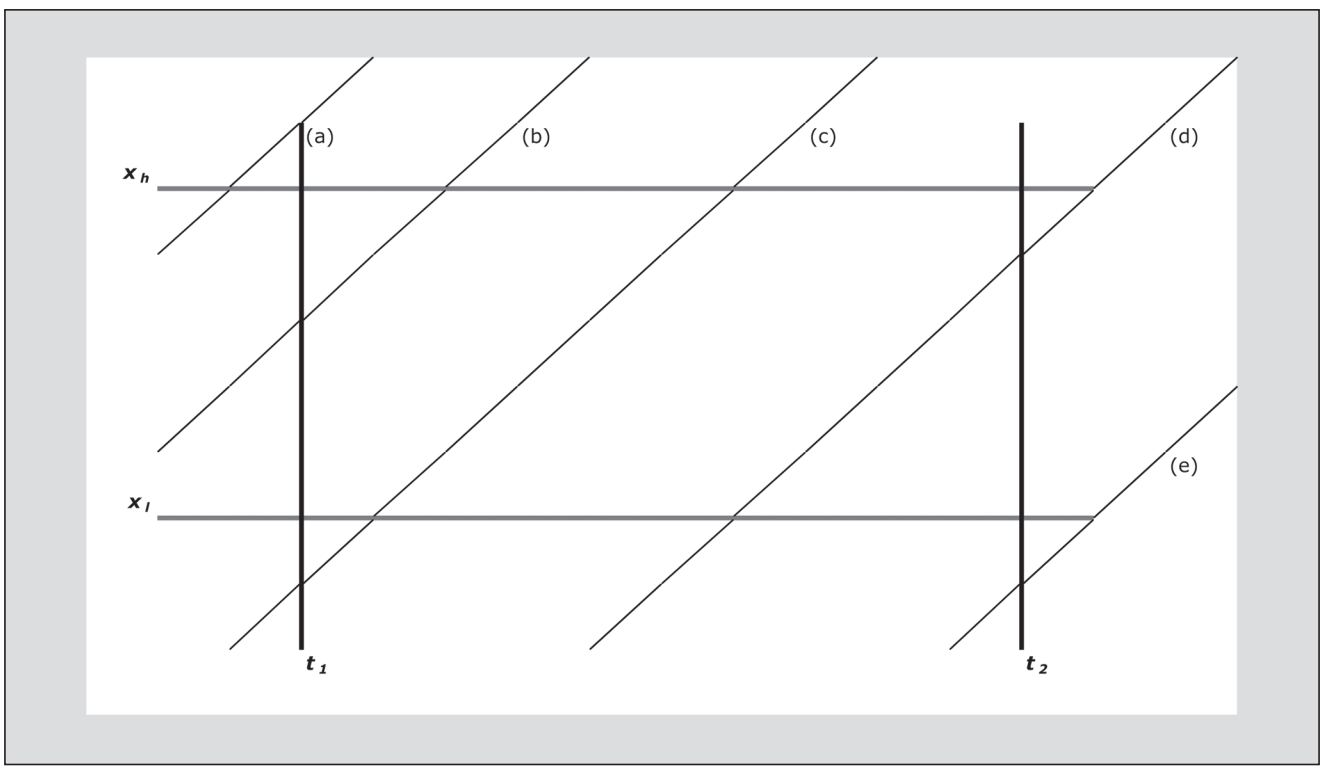
The period of investigation is the measure of the time period for which we seek to estimate mortality, and is defined as the period $(t_2 - t_1)$, where $t_2$ is the end date of the period of investigation, and $t_1$ the start date, measured in the same units as that defined by the age range.

Graphically, then, we seek to measure mortality in the age and period defined by the heavy lines in Figure 17.2.

An individual's life course by age and period is represented by the diagonal lines (as with a conventional Lexis diagram). Five possible scenarios (labelled (a) through (e)) are portrayed. Any individual's position in the space can be defined by their age at $t_1$, $x_{t1}$. It follows, further, that any person aged $x$ at $t_1$, if she or he does not die before $t_2$, would be aged $x_{t2} = x_{t1} + (t_2 - t_1)$ at time $t_2$. We define the age at death of those deaths that occur in the specified age range in the period of investigation to be $x_d$. The relative contribution of each scenario to the exposure to risk is determined by the algorithms in Table 17.2.

Applying these rules to define the exposure in the age range in the period of investigation for each individual and aggregating gives the total exposure to risk, which is the denominator for the mortality rate. Summing the deaths occurring in the age range in the period of investigation provides the numerator.

***FIGURE 17.2*** *Lexis diagram showing calculation of exposure to risk*



***Step 4: Weighting and cumulating events and exposure time***

The sample weight variable in a standard DHS recode file is *v005*. This variable has a mean of 1,000,000. To avoid the appearance of huge sample sizes (and much too narrow confidence intervals) it is recommended first to recalculate the weight as (*v005*/1,000,000). Let us call this new variable *wgt*. Mortality rates can be calculated by considering the contributions of each of the *N* children in the survey to the number of events and the total exposure time. The age-specific mortality rate age $x$ to $x+1$ (in months) in a period, $j$, is

$$M(x,j) = \frac{\sum_{i=1}^{N} D(i,x,j).wgt(i)}{\sum_{i=1}^{N} E(i,x,j).wgt(i)}$$

where $D(i,x,j)$ is a binary variable indicating the death of child $i$ at age $x$ in year $j$ (1 if the death occurs, 0 otherwise), $E(i,x,j)$ is the exposure time of child $i$ at age $x$ in year $j$, and $wgt(i)$ is the sample weight (mean 1.0) of child $i$.

***Step 5: Calculating probabilities of dying from age-specific mortality rates***

The rates calculated in Step 4 are per month of age exposure. It is therefore necessary to adapt the standard formula for deriving a period probability of dying from a rate to take this into account. Given that we have made a number of simplifying assumptions and are working with narrow age ranges, it is adequate to assume that deaths are evenly distributed across each single month age range. We can then calculate $q(x)$ as

$$q(x,j) = \frac{\frac{M(x,j)}{12}}{\left(1+\frac{M(x,j)}{24}\right)}.$$

Survivorship probabilities from birth to any age can then be obtained by chaining together survivorship by month (i.e. $(1-q(x,j))$) terms. Thus for instance

$${}_5q_0^j = 1-\prod_{x=0}^{59}\left(1-q(x,j)\right).$$

**WORKED EXAMPLE**

As noted above, direct estimation of child mortality from a birth history requires working with unit record data rather than tabulations. As a worked example, we will therefore illustrate with a limited number of records adapted from a DHS, specifically the mortality of 5-month olds in 2002 from the 2004 Malawi DHS. Only children born between 1 July 2001 and 31 July 2002 are at risk of dying at age 5 months in calendar year 2002 (children born before 1 July 2001 would be aged 6 months or more by the beginning of calendar year 2002, and those born after 31 July 2002 would not have reached age 5 months in the year). Only relevant records are shown, that is, those for births between month 1218 and 1230 in CMC terms (July 2001 to July 2002). In practice, we would also exclude any births that died before

*TABLE 17.2 Algorithm for determining exposure to risk*

| Scenario | Description | Defining rule(s) | Exposure for survivors in the period of investigation | Exposure for decedents (where death occurs in the period of investigation) |
|---|---|---|---|---|
| (a) | Aged older than $x_h$ at $t_1$ | $x_{t1} > x_h$ | 0 | 0 |
| (b) | Aged between $x_l$ and $x_h$ at $t_1$. Attains $x_h$ in the period of investigation | $x_l < x_{t1} < x_h$<br>$x_{t1} + (t_2 - t_1) > x_h$ | $x_h - x_{t1}$ | $x_d - x_{t1}$ |
| (c) | Attains $x_l$ and $x_h$ in the period of investigation | $x_l > x_{t1}$<br>$x_{t1} + (t_2 - t_1) > x_h$ | $x_h - x_l$ | $x_d - x_l$ |
| (d) | Attains $x_l$ in the period of investigation but period ends before attainment of $x_h$ | $x_l > x_{t1}$<br>$x_l < x_{t1} + (t_2 - t_1) < x_h$ | $x_{t1} + (t_2 - t_1) > x_l$ | $x_d - x_l$ |
| (e) | Does not attain $x_l$ in the period of investigation | $x_{t1} + (t_2 - t_1) < x_l$ | 0 | 0 |

five months of age, but we will include them in the example to show that we exclude them from calculations.

Table 17.3 shows the key variables for 50 records from the 2004 Malawi DHS; note that these are birth records, not woman records.

***Step 1: Manipulation of age at death and calculation of estimated birth date and age at death***

Random numbers *random1* and *random2* are derived as described above, resulting in revised values of dates of birth and age at death, *dob'* and *aad'*. The date of death *dod'* is estimated as the sum of the imputed month of birth *dob'* and imputed month of death *aad'*. Column 10 of Table 17.4 shows *dod'*.

***Step 2: Location of deaths in target year***

A relevant death in terms of period is one with a CMC between 1224 to 1235. The deaths in records 3, 6, 31 and 41 of Table 17.4 are therefore not relevant because they are deemed not to have occurred in 2002. The deaths in records 11 and 45 are not relevant because the child died at 2 months (11) or 1 month (45) of age, and therefore was not exposed to the risk of dying at age 5 months.

***Step 3: Derivation of exposure to risk***

Table 17.5 presents the calculation of the exposure to risk for the 50 cases described above. The rule used to determine the exposure is presented in the column headed 'Scenario'. The resulting exposure is presented in the following two columns for those who survive the period of investigation and those that die during the period.

*TABLE 17.3 Basic birth history data for direct estimation of child mortality*

| Record | *b3* | *b5* | *b6* | *v005* | Record | *b3* | *b5* | *b6* | *v005* |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1223 | yes | . | 469061 | 26 | 1224 | yes | . | 1095220 |
| 2 | 1223 | yes | . | 469061 | 27 | 1230 | no | 205 | 1594776 |
| 3 | 1222 | no | 107 | 469061 | 28 | 1225 | yes | . | 1594776 |
| 4 | 1224 | yes | . | 469061 | 29 | 1221 | yes | . | 1594776 |
| 5 | 1223 | yes | . | 469061 | 30 | 1225 | yes | . | 1594776 |
| 6 | 1218 | no | 205 | 469061 | 31 | 1229 | no | 208 | 1538303 |
| 7 | 1230 | yes | . | 2171218 | 32 | 1223 | yes | . | 1538303 |
| 8 | 1225 | yes | . | 704240 | 33 | 1220 | yes | . | 1538303 |
| 9 | 1230 | yes | . | 704240 | 34 | 1226 | yes | . | 1538303 |
| 10 | 1224 | yes | . | 704240 | 35 | 1225 | yes | . | 1538303 |
| 11 | 1224 | no | 202 | 704240 | 36 | 1220 | yes | . | 1538303 |
| 12 | 1221 | yes | . | 1106470 | 37 | 1224 | no | 205 | 1538303 |
| 13 | 1225 | yes | . | 1106470 | 38 | 1228 | yes | . | 1538303 |
| 14 | 1224 | no | 205 | 1106470 | 39 | 1219 | yes | . | 3789587 |
| 15 | 1221 | yes | . | 1106470 | 40 | 1228 | yes | . | 2011510 |
| 16 | 1221 | yes | . | 1106470 | 41 | 1223 | no | 302 | 2011510 |
| 17 | 1218 | no | 205 | 1106470 | 42 | 1220 | yes | . | 2011510 |
| 18 | 1229 | yes | . | 3900164 | 43 | 1220 | yes | . | 2011510 |
| 19 | 1230 | yes | . | 1247934 | 44 | 1221 | yes | . | 686252 |
| 20 | 1224 | yes | . | 1247934 | 45 | 1228 | no | 201 | 686252 |
| 21 | 1226 | no | 201 | 1247934 | 46 | 1229 | yes | . | 2451926 |
| 22 | 1221 | yes | . | 537170 | 47 | 1219 | yes | . | 2451926 |
| 23 | 1218 | yes | . | 537170 | 48 | 1219 | yes | . | 1043244 |
| 24 | 1227 | yes | . | 537170 | 49 | 1224 | yes | . | 1043244 |
| 25 | 1226 | yes | . | 537170 | 50 | 1230 | no | 205 | 1043244 |

*TABLE 17.4 Derivation of imputed date of birth, age at death and date of death, Malawi, 2004 DHS (50 cases)*

| Record | *b3* | *b5* | *b6* | *v005* | *random1* | *random2* | *dob'* | *aad'* | *dod'* |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1223 | yes | . | 469061 | 0.55 | | 1223.55 | | |
| 2 | 1223 | yes | . | 469061 | 0.85 | | 1223.85 | | |
| 3 | 1222 | no | 107 | 469061 | 0.15 | 0.05 | 1222.15 | 0.28 | 1222.43 |
| 4 | 1224 | yes | . | 469061 | 0.25 | | 1224.25 | | |
| 5 | 1223 | yes | . | 469061 | 0.25 | | 1223.25 | | |
| 6 | 1218 | no | 205 | 469061 | 0.05 | 0.45 | 1218.05 | 5.45 | 1223.5 |
| 7 | 1230 | yes | . | 2171218 | 0.55 | | 1230.55 | | |
| 8 | 1225 | yes | . | 704240 | 0.55 | | 1225.55 | | |
| 9 | 1230 | yes | . | 704240 | 0.25 | | 1230.25 | | |
| 10 | 1224 | yes | . | 704240 | 0.35 | | 1224.35 | | |
| 11 | 1224 | no | 202 | 704240 | 0.55 | 0.75 | 1224.55 | 2.75 | 1227.3 |
| 12 | 1221 | yes | . | 1106470 | 0.45 | | 1221.45 | | |
| 13 | 1225 | yes | . | 1106470 | 0.75 | | 1225.75 | | |
| 14 | 1224 | no | 205 | 1106470 | 0.85 | 0.25 | 1224.85 | 5.25 | 1230.1 |
| 15 | 1221 | yes | . | 1106470 | 0.35 | | 1221.35 | | |
| 16 | 1221 | yes | . | 1106470 | 0.45 | | 1221.45 | | |
| 17 | 1218 | no | 205 | 1106470 | 0.95 | 0.65 | 1218.95 | 5.65 | 1224.6 |
| 18 | 1229 | yes | . | 3900164 | 0.45 | | 1229.45 | | |
| 19 | 1230 | yes | . | 1247934 | 0.65 | | 1230.65 | | |
| 20 | 1224 | yes | . | 1247934 | 0.65 | | 1224.65 | | |
| 21 | 1226 | no | 201 | 1247934 | 0.75 | 0.85 | 1226.75 | 1.85 | 1228.6 |
| 22 | 1221 | yes | . | 537170 | 0.65 | | 1221.65 | | |
| 23 | 1218 | yes | . | 537170 | 0.85 | | 1218.85 | | |
| 24 | 1227 | yes | . | 537170 | 0.95 | | 1227.95 | | |
| 25 | 1226 | yes | . | 537170 | 0.85 | | 1226.85 | | |
| 26 | 1224 | yes | . | 1095220 | 0.95 | | 1224.95 | | |
| 27 | 1230 | no | 205 | 1594776 | 0.15 | 0.65 | 1230.15 | 5.65 | 1235.8 |
| 28 | 1225 | yes | . | 1594776 | 0.15 | | 1225.15 | | |
| 29 | 1221 | yes | . | 1594776 | 0.85 | | 1221.85 | | |
| 30 | 1225 | yes | . | 1594776 | 0.05 | | 1225.05 | | |
| 31 | 1229 | no | 208 | 1538303 | 0.65 | 0.85 | 1229.65 | 8.85 | 1238.5 |
| 32 | 1223 | yes | . | 1538303 | 0.45 | | 1223.45 | | |
| 33 | 1220 | yes | . | 1538303 | 0.15 | | 1220.15 | | |
| 34 | 1226 | yes | . | 1538303 | 0.55 | | 1226.55 | | |
| 35 | 1225 | yes | . | 1538303 | 0.95 | | 1225.95 | | |
| 36 | 1220 | yes | . | 1538303 | 0.45 | | 1220.45 | | |
| 37 | 1224 | no | 205 | 1538303 | 0.25 | 0.85 | 1224.25 | 5.85 | 1230.1 |
| 38 | 1228 | yes | . | 1538303 | 0.35 | | 1228.35 | | |
| 39 | 1219 | yes | . | 3789587 | 0.35 | | 1219.35 | | |
| 40 | 1228 | yes | . | 2011510 | 0.15 | | 1228.15 | | |
| 41 | 1223 | no | 302 | 2011510 | 0.65 | 0.55 | 1223.65 | 30.6 | 1254.25 |
| 42 | 1220 | yes | . | 2011510 | 0.35 | | 1220.35 | | |
| 43 | 1220 | yes | . | 2011510 | 0.25 | | 1220.25 | | |
| 44 | 1221 | yes | . | 686252 | 0.95 | | 1221.95 | | |
| 45 | 1228 | no | 201 | 686252 | 0.85 | 0.35 | 1228.85 | 1.35 | 1230.2 |
| 46 | 1229 | yes | . | 2451926 | 0.25 | | 1229.25 | | |
| 47 | 1219 | yes | . | 2451926 | 0.05 | | 1219.05 | | |
| 48 | 1219 | yes | . | 1043244 | 0.85 | | 1219.85 | | |
| 49 | 1224 | yes | . | 1043244 | 0.95 | | 1224.95 | | |
| 50 | 1230 | no | 205 | 1043244 | 0.35 | 0.35 | 1230.35 | 5.35 | 1235.7 |

For children who survive to age 6 months, those born in months 1219 to 1229 contribute a full month of exposure time to the age-period of interest (i.e. from exactly 5 to exactly 6 months). Thus record 1 (born 1223.55) contributes a full month. A child born in month 1218 will contribute *(dob*–1218) months, so record 23 (born 1218.85) contributes 0.85 of a month; and a child born in month 1230 will contribute (1231–*dob*) months, so record 7 contributes 1231–1230.55 = 0.45 months. The children born in months 1219 to 1229 who die at age 5 months will contribute (*aad*–5) months of exposure; thus the death in record 14 occurs at 5.25 months and contributes 0.25 months of exposure.

***Step 4: Weighting and cumulating events and exposure time***

The final step before calculating the death rate is to take account of the record sample weight in both the deaths and the exposure time, and then sum the weighted deaths and exposure. Columns 6 and 7 of Table 17.5 show the exposure to risk for survivors and relevant deaths. Columns 8 and 9 then multiply columns 6 and 7 respectively by the sample weight *v005*/1,000,000. The age-specific mortality rate $M(5,2002)$ is then calculated by dividing the sum of the weighted deaths by the sum of the weighted exposure time:

$$M(x,j)=\frac{\sum_{i=1}^{N} D(i,x,j)\times wgt(i)}{\sum_{i=1}^{N} E(i,x,j)\times wgt(i)}=\frac{6.389}{59.317}=0.1077\,.$$

***Step 5: Calculating probabilities of dying from age-specific mortality rates***

The rates calculated in Step 4 are per month of exposure. It is therefore necessary to adapt the standard formula for deriving a period probability of dying from a rate. Given that we have made a number of simplifying assumptions and are working with narrow age ranges, it is adequate to assume that deaths are evenly distributed across each single month age range, even for the first month of life. We can then calculate $q(x)$ as

$$q(5,2002)=\frac{\frac{M(5,2002)}{12}}{\left(1+\frac{M(5,2002)}{24}\right)}=\frac{\frac{0.1077}{12}}{\left(1+\frac{0.1077}{24}\right)}$$

$$=\frac{.008975}{1+0.004488}=0.008935.$$

Once all the $q(x,j)$ have been calculated, they can be converted into their complements, probabilities of surviving, and chained together to produce survivorship probabilities and probabilities of dying from birth to any desired age.

To obtain rates and probabilities for periods longer than a single calendar year, the weighted sums obtained in Step 4 are summed across years as required. Step 5 remains exactly the same.

Note that the procedure described here differs from that used by DHS. The DHS approach calculates probabilities directly for quasi-cohorts (Rutstein and Rojas 2003). Calculations are made for eight age groups: neonatal, 1–2 months, 3–5 months, 6–11 months, and years from age 1 to age 4. For each age range, period deaths are derived from date of birth and age at death. The risk set is an approximation of the number of children who enter that age range during the period. This approximation is the sum of all children who enter the age range and leave the age range (or would do so if they survived) during the time period, plus half of those who enter the age range during the period but would leave it after the period, plus half of those who enter the age range before the period but would leave it during the period.

Whichever procedure is used, individual-level data from the FBH will be required. Although the calculations could be carried out from detailed tables, it would be very tedious to do so. Use of a suitable computer routine is strongly recommended.

## INTERPRETATION

The key characteristic of direct child mortality estimation, namely that information is provided by surviving women who still live in surveyed households, needs to be borne in mind when interpreting results as there is risk of respondent selection bias. In particular, the mortality experience of children born in a community whose mothers no longer live in the community will not be included in the measures. If such children have higher mortality than those born to mothers who do still live in the community, mortality will be under-estimated. The most severe form of this bias is likely to result from substantial levels of HIV prevalence in the community, since such prevalence in the absence of widespread antiretroviral therapy will result in a strong positive correlation between survival of child and survival of mother (see Chapter 15). However, some positive correlation between mother and child survival is almost certain in any population. Other reasons for bias

*TABLE 17.5 Derivation of exposure to risk for estimation of child mortality, Malawi, 2004 DHS (50 cases)*

| | | | | | | Exposure to risk | | Weighted | |
|---|---|---|---|---|---|---|---|---|---|
| Record | *dob'* | *aad'* | *dod'* | *v005* | Scenario | Survivors | Deaths | Exposure | Deaths |
| 1 | 1223.55 | | | 469061 | c | 1 | | 0.469 | |
| 2 | 1223.85 | | | 469061 | c | 1 | | 0.469 | |
| 3 | 1222.15 | 0.25 | 1222.4 | 469061 | N/A | N/A | N/A | 0.000 | |
| 4 | 1224.25 | | | 469061 | c | 1 | | 0.469 | |
| 5 | 1223.25 | | | 469061 | c | 1 | | 0.469 | |
| 6 | 1218.05 | 5.45 | 1223.5 | 469061 | N/A | N/A | N/A | 0.000 | |
| 7 | 1230.55 | | | 2171218 | d | 0.45 | | 0.977 | |
| 8 | 1225.55 | | | 704240 | c | 1 | | 0.704 | |
| 9 | 1230.25 | | | 704240 | d | 0.75 | | 0.528 | |
| 10 | 1224.35 | | | 704240 | c | 1 | | 0.704 | |
| 11 | 1224.55 | 2.75 | 1227.3 | 704240 | c | 1 | | 0.704 | |
| 12 | 1221.45 | | | 1106470 | c | 1 | | 1.106 | |
| 13 | 1225.75 | | | 1106470 | c | 1 | | 1.106 | |
| 14 | 1224.85 | 5.25 | 1230.1 | 1106470 | c | | 0.25 | 0.277 | 1.106 |
| 15 | 1221.35 | | | 1106470 | c | 1 | | 1.106 | |
| 16 | 1221.45 | | | 1106470 | c | 1 | | 1.106 | |
| 17 | 1218.95 | 5.65 | 1224.6 | 1106470 | b | | 0.6 | 0.664 | 1.106 |
| 18 | 1229.45 | | | 3900164 | c | 1 | | 3.900 | |
| 19 | 1230.65 | | | 1247934 | d | 0.35 | | 0.437 | |
| 20 | 1224.65 | | | 1247934 | c | 1 | | 1.248 | |
| 21 | 1226.75 | 1.85 | 1228.6 | 1247934 | c | 1 | | 1.248 | |
| 22 | 1221.65 | | | 537170 | c | 1 | | 0.537 | |
| 23 | 1218.85 | | | 537170 | b | 0.85 | | 0.457 | |
| 24 | 1227.95 | | | 537170 | c | 1 | | 0.537 | |
| 25 | 1226.85 | | | 537170 | c | 1 | | 0.537 | |
| 26 | 1224.95 | | | 1095220 | c | 1 | | 1.095 | |
| 27 | 1230.15 | 5.65 | 1235.8 | 1594776 | d | | 0.65 | 1.037 | 1.595 |
| 28 | 1225.15 | | | 1594776 | c | 1 | | 1.595 | |
| 29 | 1221.85 | | | 1594776 | c | 1 | | 1.595 | |
| 30 | 1225.05 | | | 1594776 | c | 1 | | 1.595 | |
| 31 | 1229.65 | 8.85 | 1238.5 | 1538303 | c | 1 | | 1.538 | |
| 32 | 1223.45 | | | 1538303 | c | 1 | | 1.538 | |
| 33 | 1220.15 | | | 1538303 | c | 1 | | 1.538 | |
| 34 | 1226.55 | | | 1538303 | c | 1 | | 1.538 | |
| 35 | 1225.95 | | | 1538303 | c | 1 | | 1.538 | |
| 36 | 1220.45 | | | 1538303 | c | 1 | | 1.538 | |
| 37 | 1224.25 | 5.85 | 1230.1 | 1538303 | c | | 0.85 | 1.308 | 1.538 |
| 38 | 1228.35 | | | 1538303 | c | 1 | | 1.538 | |
| 39 | 1219.35 | | | 3789587 | c | 1 | | 3.790 | |
| 40 | 1228.15 | | | 2011510 | c | 1 | | 2.012 | |
| 41 | 1223.65 | 32.35 | 1256 | 2011510 | c | 1 | | 2.012 | |
| 42 | 1220.35 | | | 2011510 | c | 1 | | 2.012 | |
| 43 | 1220.25 | | | 2011510 | c | 1 | | 2.012 | |
| 44 | 1221.95 | | | 686252 | c | 1 | | 0.686 | |
| 45 | 1228.85 | 1.35 | 1230.2 | 686252 | c | 1 | | 0.686 | |
| 46 | 1229.25 | | | 2451926 | c | 1 | | 2.452 | |
| 47 | 1219.05 | | | 2451926 | c | 1 | | 2.452 | |
| 48 | 1219.85 | | | 1043244 | c | 1 | | 1.043 | |
| 49 | 1224.95 | | | 1043244 | c | 1 | | 1.043 | |
| 50 | 1230.35 | 5.35 | 1235.7 | 1043244 | d | | 0.35 | 0.365 | 1.043 |
| **TOTAL** | | | | | | | | **59.317** | **6.389** |

may exist. For example, high in-migration rates will result in women reporting on the survival of children born and raised elsewhere, while high out-migration will remove responses about children who were born and raised in the community. Although it is impossible to know *a priori* the direction or magnitude of such biases, the analyst needs to keep in mind their potential effect. Non-response may also be an issue if women absent from the community for an extended period cannot be interviewed in person, but may have experienced different risks to their children, or may not be present in part because their children have experienced different risks.

## EXTENSION TO THE METHOD: TRUNCATED BIRTH HISTORIES

### The truncated birth history: Data quality assessment

The truncated birth history (TBH) provides fewer opportunities for data quality checks than the full birth history (FBH) because the time series of events reported is by definition truncated. If the truncation is by time period, the events reported should be representative of the time period covered, whereas if the truncation is by number of events, the events reported may be representative only of all events in quite a short period prior to the survey, and this will complicate any assessment of the sequence of events in time.

As with the full birth history, the first step should be to examine the data for missing values. The second step should involve the examination of sex ratios at birth and heaping on ages at death.

No direct assessment of birth transference will be possible, because no detailed information about dates of births is available prior to the truncation point. However, an indirect assessment is possible. A TBH should always involve the initial collection of a summary birth history. The births and child deaths for an age group of women defined as at the survey date can therefore be calculated both at the time of the survey (from the summary birth history) and (only approximately for the deaths) at the truncation point, by subtracting the births and child deaths reported in the TBH. The calculation for births is precise, but for child deaths is approximate because some of the child deaths reported in the summary birth history (SBH) will have occurred during the post-truncation period to children born before the truncation point; typically, however, the number of such extra deaths will be small given that child mortality risks drop rapidly with age of child. The data quality assessment is therefore the comparison of the proportion dead (by age group of mother at the time of the survey) of the children born after the cut-off date to that of the children born before the cut-off date.

There are two reasons why the former proportion will generally be smaller than the latter. First, the children will have been exposed to the risk of dying for a shorter period. Second, if child mortality is falling over time, they will have been exposed to lower age-specific risks as well. However, if children who have died are systematically omitted from the post-truncation period, or if they are reported in the summary birth history but not reported as having been born in the period, the ratio of the two will be inflated by data error. We can estimate a plausible error-free ratio if data are available from a full birth history for the same population at an earlier or later date. Table 17.6 shows data from

*TABLE 17.6 Proportions of children dead by whether the birth occurred before or during the TBH date window, Mongolia, 1998, 2003 and 2008 RHS*

| Age group | RHS 1998 (FBH) | | | RHS 2003 (TBH) | | | RHS 2008 (TBH) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Proportion dead | | Ratio | Proportion dead | | Ratio | Proportion dead | | Ratio |
| | Before | After | | Before | After | | Before | After | |
| 20–24 | 0.106 | 0.070 | 1.5 | 0.222 | 0.035 | 6.3 | 0.052 | 0.041 | 1.2 |
| 25–29 | 0.140 | 0.061 | 2.3 | 0.122 | 0.036 | 3.4 | 0.083 | 0.024 | 3.5 |
| 30–34 | 0.128 | 0.082 | 1.6 | 0.117 | 0.022 | 5.4 | 0.081 | 0.015 | 5.3 |
| 35–39 | 0.072 | 0.064 | 1.1 | 0.120 | 0.025 | 4.7 | 0.097 | 0.010 | 10.2 |
| 40–44 | 0.119 | 0.068 | 1.8 | 0.150 | 0.051 | 3.0 | 0.095 | 0.010 | 9.6 |
| 45–49 | 0.213 | 0.000 | * | 0.066 | 0.048 | 1.4 | 0.119 | 0.000 | * |

Mongolia: three Reproductive Health Surveys, one in 1998 that included a full birth history and two – one in 2003 and one in 2008 – that collected only TBHs. The 1998 full birth history data are used to calculate proportions dead for children born before and after a comparably-defined cut-off date, and compared to the proportions calculated from the 2003 and 2008 TBH data. As can be seen, the TBH ratios are several times larger than the full birth history ratios, providing compelling evidence of transference of dead children out of the post-truncation period. In the absence of a country-specific baseline, such as that provided here by the 1998 RHS survey, ratios of 3 or higher should be taken as evidence of probable omission of dead children from the recent reference period.

**The truncated birth history:**
**Calculation of child mortality indicators for cohorts**

The calculation of cohort probabilities of dying from a TBH follows the same principle as that followed with a FBH: the probability of dying by age $x$ is calculated as the number of dead children to the number of children ever born in some defined cohort born no less than $x$ years before the survey. There is an important difference, however, as made clear in the Lexis Diagram in Figure 17.1, namely that the value of $x$ is constrained by the truncation date. For example, if the truncation date is 5 years before the survey, no birth cohort will have been fully exposed to the full risk of dying by age 5, and the cohorts exposed fully to risks up to age 2 are limited to births 2, 3 and 4 years before the survey. Thus there are limits to the range of ages for which mortality indicators can be derived.

**The truncated birth history:**
**Calculation of child mortality indicators for time periods**

The basic approach to calculating standard indicators from a TBH follows the same principles as that used for a full birth history: to calculate age-specific rates for a specified time period, convert them into estimates of probabilities of dying in successive age intervals, and apply the probabilities to a synthetic cohort of births to create the life table. The problem with analysing a TBH in this way is the same, however, as that faced in calculating cohort indicators, namely that cases and exposure time become progressively more restricted as age increases. Thus if the cut-off point is five years before the survey, the measures for ages 3 and 4 will be based on small numbers and have wide sampling errors.

## REFERENCES


Preston SH, P Heuveline and M Guillot. 2001. *Demography: Measuring and Modelling Population Processes.* Oxford: Blackwell.

Rutstein S and G Rojas. 2003. *Guide to DHS Statistics*. Calverton, MD: ORC Macro.

# Chapter 18 Childhood mortality estimated from health facility data: The Preceding Birth Technique

*Allan G Hill*

In the introduction to child mortality (Chapter 15), we have already drawn attention to the possibility of using data collected at health facilities to measure early childhood mortality. Setting selection issues aside for the moment, the most useful technique for obtaining such mortality measures is based on a simple question put to a mother when expecting or delivering her next child, here referred to as the index child. At the time of the pregnancy or delivery, the supplementary information needed is whether or not the previous live birth is alive or dead at the time of the subsequent pregnancy or birth of the index child. For a set of mothers (generally not less than 1,000 respondents), the proportions dead amongst the previous live-born children are then converted into a measure of early childhood mortality. This measure is usually close to ${}_2q_0$, but can be closer to ${}_3q_0$ when birth intervals are long. Variations of the method also allow the proportions of preceding children dead obtained before the delivery of the index child, for example at the time of an antenatal visit, or after the delivery of the index child, to be converted into measures of early childhood survival.

There are several attractive features of child survival information gathered in this way from hospitals, clinics and other health centres. First, the information often forms part of the routine health system reporting so the need for special studies and surveys is obviated. Second, when such information comes from health facilities, additional information which is difficult to obtain in retrospective surveys can be obtained relatively easily. This includes characteristics of the mother as well as key attributes of the births including sex and birth order, and birth weight. Third, the data can be disaggregated to provide detailed estimates for particular health facilities (when the population in the clinic's catchment area is sufficiently large), for towns and small provinces. Such local or facility-based information could be useful to health authorities intent on targeting the communities with the worst infant and child mortality rates. Trends at the local level can also be used to assess the effectiveness of past health interventions. Finally, we can expect the data in health facilities to be reasonably accurate as they are mostly being collected by literate professionals. Further, clinic-based respondents may be more prepared than mothers to report events which are otherwise seen as stigmatizing (distinguishing live births, still births, abortions and miscarriages) or painful to recall (a neonatal or infant death).

## ORIGIN OF THE PRECEDING BIRTH TECHNIQUE METHODS

The technique has its origins in a study of mortality in the Solomon Islands in the 1980s. In the course of this study it was noted that amongst the information routinely collected in maternity centres were answers to the questions on children ever-borne and surviving as well as a question on the survival of the preceding born child if the mother was delivering her second or subsequent child, the index birth (see Figure 18.1). Brass and Macrae set out to ascertain how these data could be related to conventional measures of child survival estimated from summary birth histories and the proportions dead of preceding births (Brass and Macrae 1984, 1985). Two methods were proposed. The Preceding Birth Technique (PBT) has attracted most attention since it provides a running estimate of early childhood mortality close to the current period. The theoretical basis of the method has been expanded and developed to allow applications to data on the survival of preceding children collected before and after a birth (Aguirre 1994; Aguirre and Hill 1988; Hill and Aguirre 1990). The second method (Brass and Macrae 1985), based on the total numbers of children

ever-borne and surviving requires more refinement, is more complicated to apply and is not discussed further here.

Subsequently, others have applied the PBT in a variety of circumstances for different purposes – for example, in refugee camps, to measure abortion rates in antenatal clinics, for small area estimation and to measure the impact of health interventions (Bicego, Augustin, Musgrave *et al.* 1989; Madi 2000; Oliveras, Ahiadeke, Adanu *et al.* 2008; Rowe, Onikpo, Lama *et al.* 2011). Research in a demographic surveillance site where births and deaths had been accurately and independently recorded showed that good results could be obtained when contraception had become widespread and birth intervals had lengthened as well as when the data had been collected at antenatal rather than maternity clinics or at the time of the first vaccination of the new born (Bairagi, Shuaib and Hill 1997).

## CAVEATS AND WARNINGS

The analyst faces several difficulties in making use of facility-based data on child mortality. The most important of these is that the population attending the facilities is not randomly selected. The resulting selection biases are important in three main ways.

The first bias in facility-based data arises because of the incomplete coverage of the population by health facilities, both public and private. This selection can work in different ways. Often, the urban population has easier access to health facilities than the rural population. The better off and better educated often make greater use of modern health services than the poor and the illiterate. This bias would likely result in estimates that are too optimistic in respect of infant and child mortality. In some cases, however, the bias can work in the other direction. The tertiary referral centres, which are generally the central maternity or teaching hospitals, often have much worse outcomes than peripheral centres simply because most complicated cases requiring surgery and other forms of advanced care are referred to these centres. Estimates based on such facilities will therefore tend to be over-estimate infant and child mortality.

The extent of this first bias is reduced when the coverage rates of health facilities used as a source of data are high. Even in sub-Saharan Africa, using the most recent DHS surveys for 38 countries, we find that 51 per cent of mothers delivered in a health facility, 76.5 per cent of mothers in urban areas. For antenatal visits, an even higher proportion, 93 per cent, were seen by a doctor or a health professional in urban areas in sub-Saharan Africa with a surprising 76 per cent being seen for antenatal care by a doctor or health professional even in rural areas (Macro International Inc 2012). With these high coverage figures, it is possible to address some of the biases associated with incomplete coverage of the population. We therefore present below a method to estimate early child mortality from the proportions of preceding born children dead when mothers attend antenatal clinics or even vaccination clinics. These options are discussed below in the section entitled '*Extensions*'.

When coverage of the population by the health facilities is much lower, selection bias is clearly more important. Survey or census data for the whole population may be used to adjust the figures coming from health facilities by comparison of the characteristics of users and non-users of the health services. An added complication is that many countries are trying hard to extend the coverage of their health services. This may add new sub-populations with distinctive mortality patterns to the pool of information on child survival. Such changes in coverage can make the interpretation of trends over time difficult. In most large populations, however, new facilities take time to add and thus the coverage of existing facilities changes quite slowly.

The second source of systematic bias is that virtually all the women seen in maternity hospitals and health centres are attending these facilities because they are about to have a baby. Any information gathered from these women is specific to these moments in their reproductive careers. By contrast, in random sample surveys, women are interviewed without any reference to the current stage of their reproductive life and so the information obtained from them is representative of the reports from all, or all parous, women. An adjustment has been proposed to make the facility-based reports on total children ever-born and dead more like the reports in household surveys. The adjustment, however, now seems too dependent on various assumptions such as the effect of birth order on child survival and the location in time of child deaths (Brass and Macrae 1985).

A third selection bias that arises with this method is the fact that each woman's last birth is never reported because there is no subsequent 'index' birth. This bias is probably trivial in high fertility settings, but in a population in which many women have only two children, there will be an over-representation of first births, which typically have above average mortality. In a population in which a substantial proportion of women have only one birth, the preceding

birth technique will also not provide unbiased estimates of early childhood mortality.

## DETAILED DESCRIPTION OF THE METHOD

A diagram helps to understand the terms used in this explanation. Figure 18.1 shows an idealized birth interval $\bar{I}$ for a woman with the three possible points of contact with the health services – for an antenatal visit, at the time of the delivery, and after the delivery for a post-partum check-up or vaccination of the new born infant.

Models and empirical data have shown that the proportions dead amongst preceding birth children, $Q$, collected close to the time of the delivery of the index birth are close to the probability of dying by an age which is close to 80 per cent of the mean live birth interval, $\bar{I}$. Tabulation of the median birth intervals for 35 of the most recent surveys covering the period 1990–2010 in sub-Saharan African countries shows that the median birth interval was 34.8 months (Macro International Inc 2012). The only countries with median birth intervals over 40 months were Ghana (40), Namibia (42), South Africa (47) and Zimbabwe (47). The proportions of preceding births dead at the time of delivery of the index child is thus close to the proportions dead by the second birthday, in life table notation, ${}_2q_0$.

The reason that the proportion of previously born children who have died by the time of a subsequent birth closely approximates ${}_2q_0$ in a life table is because the proportion dead is the integrated product of two asymmetrical functions. One function is the distribution of births over time before the most recent birth, $b(x)$. The second function is the cumulative probability of dying, $q(x)$, taken from the early part of a life table. The monthly distribution of previous births is skewed, with no live births occurring during the nine months preceding the current maternity but with a concentration of births around the mean birth interval, and a long tail stretching back in time before the most recent birth. The cumulative probabilities of dying in childhood in any life table rise quickly during the first two years of life but thereafter the cumulative proportions dead, $q_x$, flatten out beyond the age of two years (Hill and Aguirre 1990).

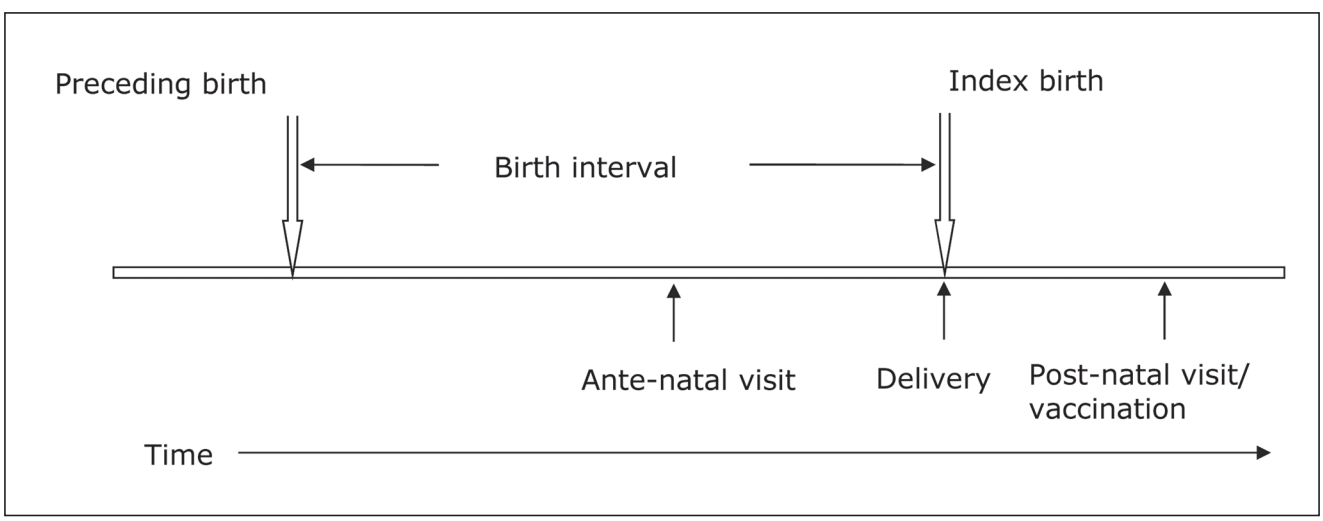
Figure 18.2 illustrates the shape of these two functions using real birth interval data (hence the slight irregularities attributable to date misreporting) and the probabilities of dying by month since birth taken from the UN General Model life table with a life expectancy at birth of 60 years.

We see that the proportion of previously born children who have died, $Q$, is thus the integrated product of these two functions. Mathematically,

$$Q = \int_0^\infty b(x)q(x)dx \qquad (1)$$

where $b(x)$ is the number of births which occurred $x$ months before the current maternity, and $q(x)$ is the cumulative probability of dying by age $x$.

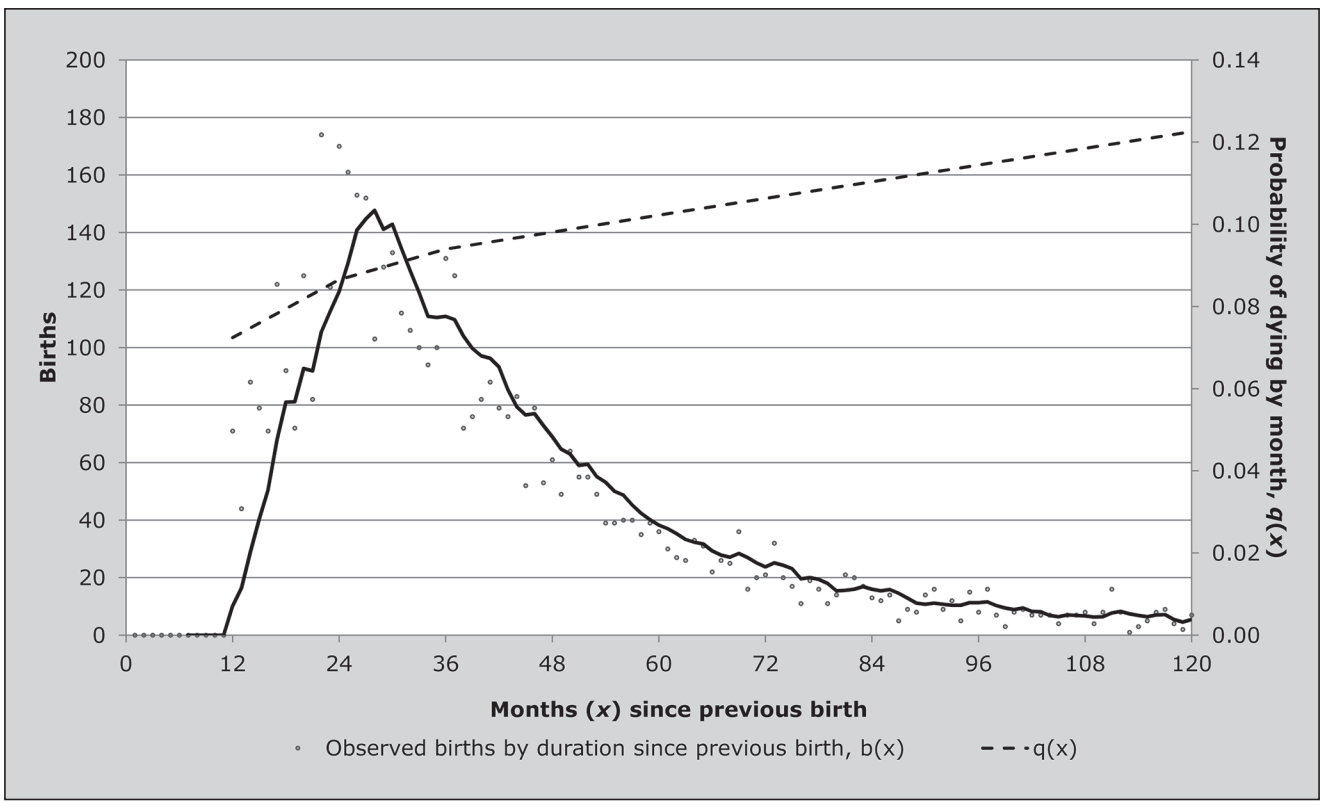
*FIGURE 18.1 An idealized birth interval with possible points of contact with the health services*



Since the method is primarily designed to measure year-to-year changes in child survival by facility, the main interest will be in the value of the index rather than in its exact representation in the life table. The main confounding effect in comparing results across different populations or over time will result from differences in birth interval length. As Rutstein's analysis has shown, however, median birth intervals have changed very slowly between the first and last DHS surveys in each country and especially in sub-Saharan Africa, so temporal effect changes in birth interval length are probably of minor importance (see Rutstein (2011: Table 2.2a)). The convention is that the proportions of preceding children dead collected at the time of a subsequent birth are simply referred to as the "index of early childhood mortality", and taken as a close approximation to ${}_2q_0$ in most populations and to ${}_{2.7}q_0$ in populations with birth intervals closer to 40 months. Clearly, many low fertility countries have longer birth intervals but most such countries have good vital registration systems and so will not be the main users of the PBT.

We can estimate the possible effects of birth intervals differing from 30 months on the child mortality measures by using model life tables. We use the UN General Standard model for both sexes combined with a life expectancy at birth of 60 years to calculate the monthly probabilities of dying up to age 5. We then calculate the percentage differences in the measure of early childhood mortality when the birth interval differs from 30 months and the exposure time is not $0.8 * I$ or 24 months (${}_2q_0$). When the birth interval is 25 months and hence the exposure time is $25 * 0.8 = 20$ months, as might occur if the data are collected at antenatal visits, then the proportions of preceding children dead is closer to ${}_{1.7}q_0$ or 5 per cent lower than if the birth interval had been 30 months. With birth intervals as long as 40 months, the proportions of preceding birth children dead approximates to ${}_{2.7}q_0$, a 7 per cent difference from the central value of ${}_2q_0$ associated with a birth interval of 30 months. If the interval between the birth of the preceding birth and the time the data are collected is as long

*FIGURE 18.2 Typical distributions of birth interval length by time since preceding birth and the cumulative risks of dying by age (x)*



as 45 months, as might occur if the mothers are seen some time after the birth of the index child, then the child mortality measure estimated is approximately ${}_3q_0$ or 9 per cent more than ${}_2q_0$ in the model life table. Although important, these differences are not large and the percentages are likely to remain the same in the short term.

Model life tables can be used to interpolate between the various measures of child mortality derived from the PBT. With birth intervals of 30 months, we are estimating approximately ${}_2q_0$. By using logit transformations of model life tables, however, we can readily derive corresponding values of ${}_1q_0$ (infant mortality) and ${}_5q_0$ (the U5MR used by UNICEF). An example of the interpolation method is shown in Table 18.2. The same procedures can be used to derive ${}_2q_0$ when the intervals between the preceding birth and the index birth are not 30 months if the birth interval is curtailed by collection of the data at antenatal visits or if the time since the birth of the preceding birth is extended by collection of the data say, at first vaccination of the index child. The associated worksheet (see website) shows, in detail, how to do this.

Numerous questions arise from the simple result that $Q \approx {}_{(\bar{I}.0.8)}q_0$. First is the issue of the omission of women with only one birth and thus no preceding birth. In most populations lacking full vital registration, however, most women proceed to have at least a second child so the mortality experience of first births is not omitted from the data and consequently this bias is small. There may, however, still be a concern in low fertility populations that first births are over-represented in the data. Second, women who die in childbirth may not survive long enough to report on the survival of their previous children, although clinical records are often available ahead of the death. We know that the risks of losing subsequent children are strongly associated with a maternal death, as the Bamako data bear out (Hill and Aguirre 1990). Fortunately maternal deaths are sufficiently rare as to have only a small effect on the data collected this way.

It is important to establish the time reference of the PBT estimates. Again Figure 18.2 helps us to estimate the mean time at death of preceding children who died before the index birth. The combination of relatively high risks of dying early in life (see the $q(x)$ function) combined with the concentration of births around the mean birth interval points to a mean age at death substantially less than half the birth interval. From models, Aguirre (1990) showed that the mean time location of deaths of the preceding born children was about two-thirds of the birth interval length before the date of birth of the index children. From empirical data with a wide range of birth intervals, the range was between 54 per cent and 74 per cent of the birth interval in months before the dates of birth of the index children. In most applications, it is recommended that analysts take the time reference to be two-thirds of the preceding birth interval before the birth of the index children. This assumption has been built into the accompanying estimation spreadsheets (see website).

**Data requirements**

The key sequence of essential questions for this technique is simple, assuming the woman being interviewed in the health facility is pregnant (antenatal visit), newly delivered (in a maternity clinic or hospital) or has brought a young infant for immunization:

- *'Were you pregnant another time before this current pregnancy/birth?'*
  If 'yes', continue. If 'no', stop.
- *'What was the result of this previous pregnancy?'* (live birth, still birth, miscarriage or abortion – spontaneous or induced).
  If 'live birth' continue. If other, stop.
- *'Is this previous born child still alive today?'* (yes/no)

Date of interview (usually date of the delivery of the index child) is also needed but this generally forms part of the standard administrative records.

The above are the basic questions. Others related to the care of the mother and her children are often added, such as the date of birth of the preceding child, its sex, its birth weight (if known), whether a singleton or multiple birth, whether still being fully breastfed and so on, depending on needs and circumstances. Similarly, some additional information on the mother (such as age, education residence) as well as information on the current delivery such as birth weight and place and type of delivery (normal vaginal, forceps, vacuum, caesarean etc.) can be relevant for maternal and child health care (if the mother is seen post-partum). Collecting the date of the birth of the child preceding the current delivery provides useful information on the average birth intervals in the population under study.

Often, the data are obtained in clinics in the form of registers or ledgers. An example of one is shown in Table 18.1 below. The content of each of the columns can be varied for different purposes but the key questions for the estimation of early child mortality are clearly the questions in columns

(7) and (8). Note that in health systems gravidity (total number of pregnancies, however short their duration) often replaces the total number of live births but with training, health workers can readily distinguish the more medical definitions (gravidity, parity, confinements) from the more demographic terms (pregnancies, live births and living children).

## WORKED EXAMPLE

The basic form of the analysis is very simple. It consists of dividing the number of preceding children dead (Row B in Table 18.2) as at the time of the index birth by the total number of preceding live-born children. Still births are excluded from the calculations.

In cases where there is an interest in estimating other life table measures, infant and under-5 mortality can be estimated using a standard from model life tables and logit transformations. In Table 18.3, we illustrate the steps involved in using logit transformations of model life tables values to produce values of ${}_1q_0$ and ${}_5q_0$. These methods are included in the associated spreadsheets (see website).

In some circumstances, information on the survival of the second-to-last born children, the child born before the preceding birth (if any), is also collected and utilized. In crude terms, the period of exposure to the risks of dying for this second-to-last born child will be slightly shorter than twice the mean birth interval, $\bar{I}$. Taking $\bar{I} = 30$ months, the proportions of second-to-last children dead at the time of the current maternity will thus be approximately ${}_{2.\bar{I}}q_0$, i.e. ${}_5q_0$. The reported proportion of second-to-last children who have died is close to the probability of dying during the first five years of life and not some younger age in this case, as the monthly birth distribution and the cumulative probabilities of dying are much flatter around the age of five than around the age of two years. The difficulty is that these data on the second-last child can be obtained only from mothers who have had at least three deliveries or at least two deliveries and a third pregnancy. Thus, the systematic selection of women with higher parities and probably with higher fecundity (shorter birth intervals) exacerbates biases relative to the case of the simple preceding birth version of the method (Hill and Aguirre 1990). Although included in the illustration below, the use of information on the survival of second-to-last born children to estimate recent child mortality is not recommended.

In Figure 18.3, we show the relationship of the different mortality measures estimated in Table 18.3. The data at face value suggest that child survival was improving in the period

*TABLE 18.1 Example of data collection register for implementation of the preceding birth technique*

| Date of delivery (1) | Mother's name or ID (2) | Mother's age or date of birth (3) | Gravidity (total pregnancies) (4) | Live births (5) | Living children (6) | This delivery: type (live birth, still birth, abortion or miscarriage) (7) | Singleton or multiple birth? (8) | If live birth preceding birth alive today? (9) | Sex of last delivery: male/ female (10) |
|---|---|---|---|---|---|---|---|---|---|
| 27 Jan 2012 | Mariama Sow | 31 Oct 1980 | 7 | 5 | 4 | Live birth | Single | Yes | M |
| 28 Jan 2012 | Comfort Frempong | 27 June 1991 | 3 | 3 | 2 | Still born | N.A. | N.A. | N.A. |
| 29 Jan 2012 | Huda Khalaf | 19 Oct 1992 | 3 | 2 | 2 | Live birth | Twin | 1 – Yes | M |
| 29 Jan 2012 | Huda Khalaf | 19 Oct 1992 | 3 | 2 | 2 | Live birth | Twin | 2 – No | F |
| 30 Jan 2012 | Mary Kenyatta | 22 yrs | 3 | 2 | 1 | Miscarriage | N.A. | N.A. | N.A. |

Note: In this example, words have been used instead of codes but in most applications, the optional answers would be pre-coded to standardize responses and to minimize the work of the health staff. The summary counts of pregnancies, live births and living children exclude the most recent pregnancy or birth. Twins or triplets need to be separately recorded – see table.

before the data were collected in Bamako's maternity clinics. We must remember, however, the selection effects implicit in using the data from second-to-last births since only women with three or more births provide the information used to estimate the measures in the last row of Table 18.3.

*TABLE 18.2 Preceding birth technique estimates of early childhood mortality for Bamako, Mali in 1985*

| Measures | Preceding births | Second-to-last births |
|---|---|---|
| Total alive amongst preceding live births (A) | 4778 | 3737 |
| Total dead amongst preceding live births (B) | 679 | 620 |
| Proportions of preceding born children dead (B /A) | 0.142 | 0.166 |

Source: Data from the Bamako maternity clinics study (Hill and Aguirre 1990)

**Extensions of the method**

A common criticism of the PBT method is that the rates are necessarily calculated for the population attending the health centres and maternity clinics and tell us nothing about the child mortality rates amongst mothers not attending such centres. As noted above, in most countries more and more mothers are giving birth in health centres of one kind or another so that gradually, the PBT estimates of child mortality will become more representative. In the interim, it is often worth exploring collection of the essential information elsewhere.

The most obvious opportunity to contact a larger proportion of mothers arises at the moment of first antenatal booking. As noted above, the proportion of pregnant women who attend such clinics is now quite high even in places with very low levels of vital registration (and delivery in health facilities). Very few adjustments are needed to adapt the collection of the essential information for the PBT in antenatal clinics. The main risk is that the information is

*FIGURE 18.3 The relationship between the proportions of preceding born children dead amongst preceding and second-to-last born births, Q, and life table measures of child mortality, ${}_1q_0$ and ${}_5q_0$*

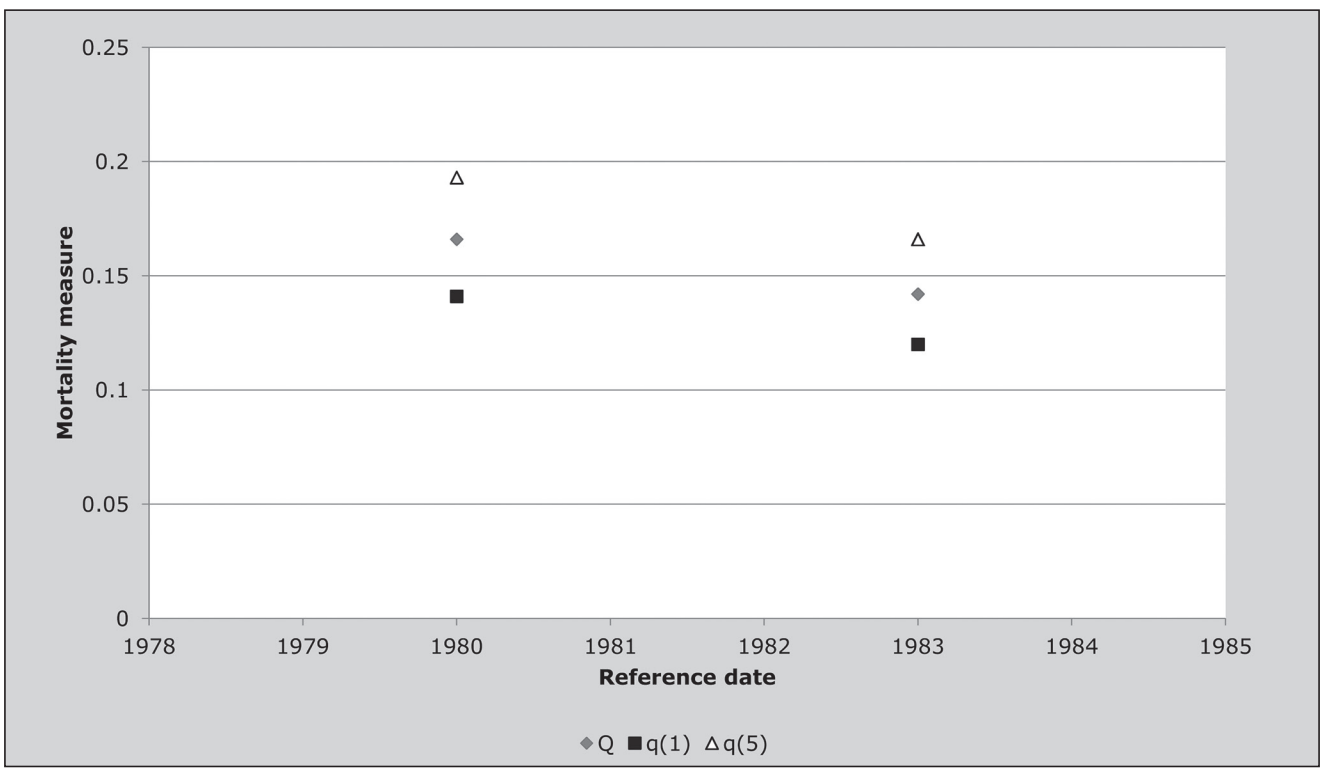


*TABLE 18.3 Using logits and the UN General Standard model life table to estimate values of infant and under-5 mortality from the proportions of preceding and second-to-last births dead at the time of a subsequent delivery*

| Measure | Proportion of preceding births dead | Proportions of second-to-last births dead |
|---|---|---|
| Observed proportions dead | 0.142 | 0.166 |
| Proportions alive | 0.858 | 0.834 |
| $l(2)$ in UN General model life table: $e(0)=60$ | 0.914 | 0.914 |
| logit $l(2)$ | –1.179 | –1.179 |
| logit $l(2)$ observed | –0.899 | –0.807 |
| Alpha | 0.28 | 0.372 |
| ${}_1q_0$: estimated infant mortality | 0.12 | 0.141 |
| ${}_5q_0$: estimated under 5 mortality | 0.166 | 0.193 |

collected multiple times from the same mother. Care must therefore be taken to ensure that the preceding birth technique information is obtained at first booking and not at all subsequent visits.

The main technical issue to be resolved is that compared with the time of delivery, the preceding birth interval will be curtailed when the information is obtained during an antenatal visit. This means that the proportion dead amongst preceding births will probably be a slight underestimate of the true value of ${}_2q_0$ in the population. In many instances, however, pregnant women do not present for the first antenatal visit until the pregnancy is well advanced. In these cases, the curtailment in the length of the birth interval will be a matter of a few months. The slight reduction in exposure has only a very small effect on the measure of early childhood mortality (Bairagi, Shuaib and Hill 1997; Hill and Aguirre 1990).

Another alternative for obtaining the key information to apply the Preceding Birth Technique method is to include the questions at the time of the first vaccination of the baby since coverage rates for vaccination are often quite high both for mothers giving birth in clinics and for mothers who give birth at home. Again the basic format of the questions remains the same. Careful wording and training is needed to distinguish the new-born child being vaccinated from the preceding birth whose survival has to be established to use the preceding birth technique. Since the vaccination protocol begins – or should begin – in the first month of the new-born baby's life, the extension of the exposure time of the preceding child to the risks of dying compared to the time of delivery is very short. The proportions of preceding born children dead at the time of vaccination can therefore be taken as a good approximation of ${}_2q_0$. There may be problems in preventing multiple reporting from mothers who return for several rounds of vaccination but this can be managed (Hill and Kelly 1996). More important is the effect of mortality of the new-born child since mothers who lose a child soon after birth will not appear in vaccination clinics. If mortality of successive children is not independent, this would lead to an overall under-estimate of the index of early child mortality.

### *Trends over time and comparisons between areas*

One important application of the preceding birth technique is for the study of early childhood mortality trends in sub-populations followed over time or for the comparison of child mortality trends within different sub-populations. There are several countries, ranging from Senegal and Mali to the Sultanate of Oman, where ongoing estimates of early childhood mortality are derived from information collected at maternity clinics. The example below is derived from the Matlab Thana surveillance site in Bangladesh where we have the added advantage of well-recorded vital registration data alongside the information on the survival of the previously born child collected at the time of the birth. In the Matlab Thana study area it is, in addition, possible to compare the childhood mortality in the 'Treatment' and 'Comparison' areas, providing an additional test of the validity of the information from the method.

In Figure 18.4 below the middle lines in the Treatment and the Comparison areas illustrate the trend in early childhood mortality derived from the questions on the survival of the preceding birth collected at the time of the birth of the index child. The lowest lines show the time trend in infant mortality derived from the vital registration system while the upper lines are ${}_3q_0$, also calculated from the vital registration data. The measure of early childhood mortality matches very closely the trend in ${}_3q_0$ since in this population, where birth intervals are close to 40 months, the Preceding Birth Technique measures mortality to about age 2.7 years or ${}_{2.7}q_0$. The goodness of fit between the directly measured child mortality measures and the proportions of preceding children dead is encouraging.

***Comparison of mortality trends in sub-populations***

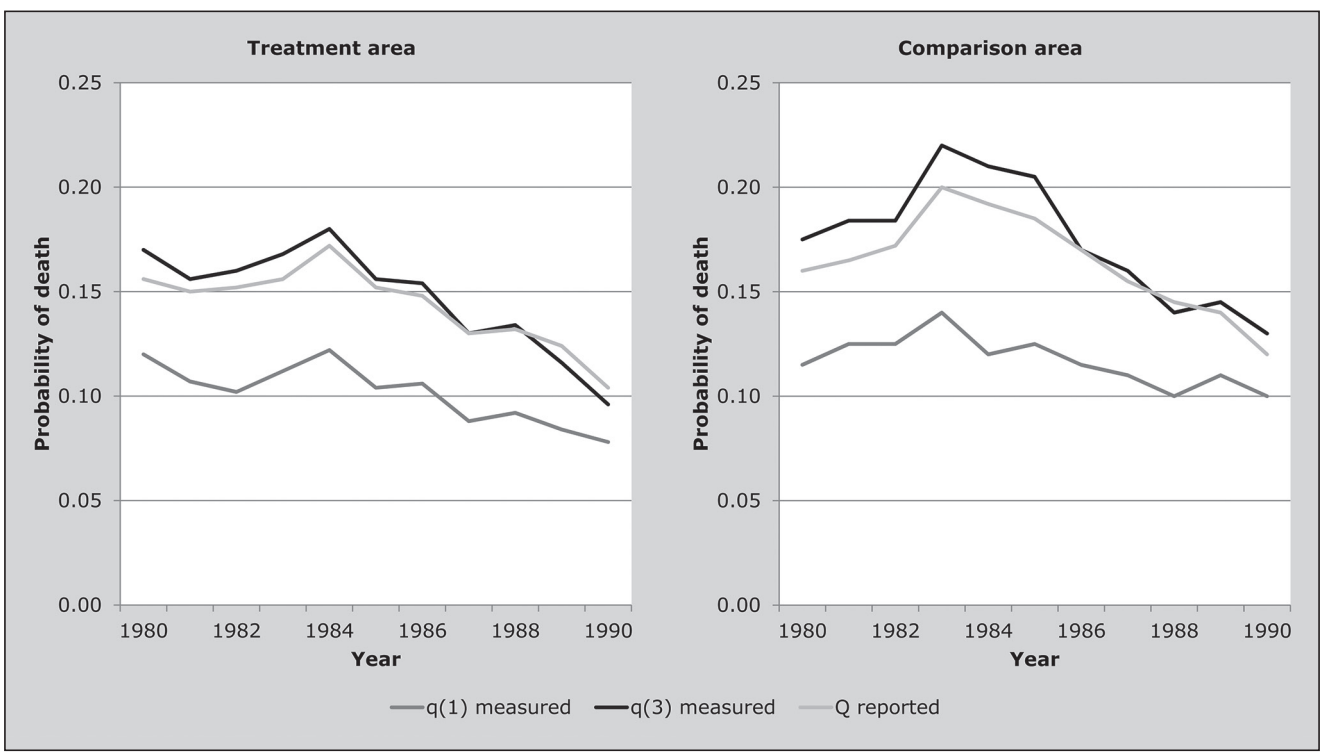
The final illustration comes from the study of early childhood mortality in the maternity clinics of Bamako, Mali where information on the birth weight of the last born child (the infant delivered in the maternity clinic) was also recorded. As Table 18.4 illustrates, there is a very strong relationship between the birth weight of the index child, the most recent birth, and the survival of the preceding born children. This analysis clearly illustrates the concentration of excess risks in certain mothers and makes the case for targeting such high-risk women in order to reduce early childhood mortality. In the same study, there were enough births occurring in each maternity facility over the course of a year to be able to calculate the index of early childhood mortality for each facility. This formed the basis of a rank ordering of facilities according to the proportions of preceding children dead at the time of subsequent delivery and thus allowed the identification of underperforming facilities, together with populations in their catchment areas, which were in need of extra resources.

*TABLE 18.4 Proportions of last and second-to-last births dead by the time of a subsequent birth, by the birth weight of the most recent born child*

| Birth weight of the index child (grams) | Preceding births | |
|---|---|---|
| | *N* | Proportions dead ≈ $_2q_0$ |
| 1500–1999 | 76 | 0.197 |
| 2000–2499 | 409 | 0.161 |
| 2500–2999 | 1389 | 0.153 |
| 3000–3499 | 1827 | 0.136 |
| 3500–3999 | 607 | 0.104 |
| 4000 or more | 98 | 0.092 |

Source: Bamako maternity clinics study (Hill and Aguirre 1990)

*FIGURE 18.4 Proportions of preceding children dead at the time of a subsequent delivery, Q, compared with infant mortality, $_1q_0$ and mortality before the third birthday, $_3q_0$, measured directly from the Matlab, Bangladesh surveillance data*



## REFERENCES


Aguirre A. 1990. "The Preceding Birth Technique for the Estimation of Childhood Mortality: Theory, Extensions and Applications." Unpublished PhD thesis, London: University of London.

Aguirre A. 1994. "Extension of the preceding birth technique", *Genus* **50**(3–4):151–169.

Aguirre A and AG Hill. 1988. "Estimacion de la mortalidad de la ninez mediante la tecnica del hijo previo con datos provenientes de centros de salud o de encuestas de hogares: aspectos metodologicos. (Estimating child mortality using the previous child technique, with data from health centers and household surveys: methodological aspects)", *Notas Poblacion* **16**(46–47):9–39.

Bairagi R, M Shuaib and AG Hill. 1997. "Estimating childhood mortality trends from routine data: a simulation using the preceding birth technique in Bangladesh", *Demography* **34**(3):411–420. doi: 10.2307/3038293

Bicego G, A Augustin, S Musgrave, J Allman and P Kelly. 1989. "Evaluation of a simplified method for estimation of early childhood mortality in small populations", *International Journal of Epidemiology* **18**(4 Suppl 2):S20–32. doi: 10.1093/ije/18.Supplement_2.S20

Brass W and S Macrae. 1984. "Childhood mortality estimated from reports on previous births given by mothers at the time of a maternity: I. Preceding-births technique", *Asian and Pacific Census Forum* **11**(2):5–8. http://hdl.handle.net/10125/3561

Brass W and S Macrae. 1985. "Childhood mortality estimated from reports on previous births given by mothers at the time of a maternity: II. Adapted multiplying factor technique", *Asian and Pacific Census Forum* **11**(4):5–9. http://hdl.handle.net/10125/3562

Hill AG and A Aguirre. 1990. "Childhood mortality estimates using the preceding birth technique – some applications and extensions", *Population Studies* **44**(2):317–340. doi: 10.1080/0032472031000144616

Hill AG and PG Kelly. 1996. *Sur la mise en place de la technique de l'accouchement precedent, Senegal, 19 au 26 Janvier 1996 (On the implementation of the Preceding Birth Technique, Senegal 19–26 January 1996)*. Arlington,VA: Partnership for Child Health Care.

Macro International Inc. MEASURE DHS STATcompiler. http://www.measuredhs.com.

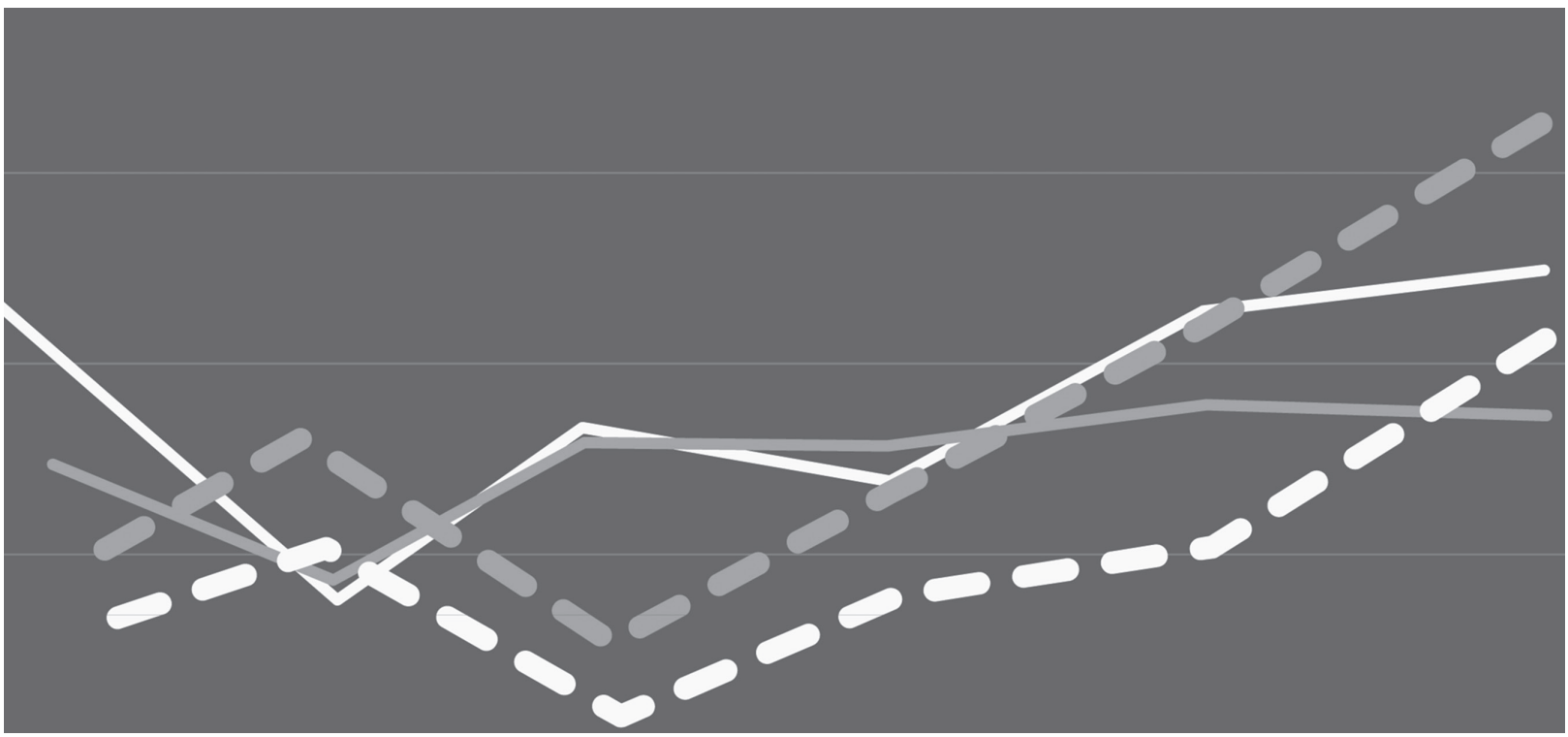
Madi HH. 2000. "Infant and child mortality rates among Palestinian refugee populations", *Lancet* **356**(9226):312. doi: 10.1016/S0140-6736(00)02511-3

Oliveras E, C Ahiadeke, RM Adanu and AG Hill. 2008. "Clinic-based surveillance of adverse pregnancy outcomes to identify induced abortions in Accra, Ghana", *Studies in Family Planning* **39**(2):133–140. doi: 10.1111/j.1728-4465.2008.00160.x

Rowe AK, F Onikpo, M Lama, DM Osterholt and MS Deming. 2011. "Impact of a malaria-control project in Benin that included the integrated management of childhood illness strategy", *American Journal of Public Health* **101**(12):2333–2341. doi: 10.2105/AJPH.2010.300068

Rutstein S. 2011. *Trends in Birth Spacing*. DHS Comparative Reports No. 28. Calverton, MD: ICF Macro. http://www.measuredhs.com/pubs/pdf/CR28/CR28.pdf

# Adult Mortality



# Chapter 19 Introduction to adult mortality analysis

*Ian M Timæus, Rob Dorrington and Kenneth Hill*

Accurate knowledge of adult mortality levels and trends in the developing world is hampered by a widespread lack of complete vital registration systems. Although knowledge of infant and child mortality once faced similar barriers, survey-based techniques – indirect methods and birth histories – have been more successful at measuring child than adult mortality, and we know correspondingly less about the latter than the former.

For the purposes of demographic analysis, adult mortality is usually defined as mortality at ages 15 or more. In some contexts though, the term 'adult mortality' is used to refer solely to mortality between exact ages 15 and 60, and is contrasted with older-age mortality, which is used to refer to mortality at ages 60 or more. The probability that a person on their 15th birthday dies before their 60th birthday, (${}_{45}q_{15}$ in the life table) has become a widely used indicator of adult mortality defined in this more restricted way.

In countries that lack complete vital registration systems, the sources of data and methods that are used to study mortality in adulthood usually differ from those used to study mortality in childhood. Some of the methods for adults can be extended to study the mortality of children aged 5 or more, but none of them are reliable sources of information on under-five mortality.

Several general issues make the study of adult mortality inherently more challenging than that of children. First, in broad terms, adult mortality rates for much of the age range are an order of magnitude lower than those of children. Adult deaths are relatively rare events. Obtaining precise measures of adult mortality requires data either on a large sample of people or on events occurring during a long reference period. Second, it is difficult to identify an appropriate informant who can provide reliable information about deceased adults. Data on child mortality can usually be collected from mothers. In addition, the characteristics of parents are among the more important determinants of the risk of dying in childhood. Since there is no single universally-suitable informant to provide data about adult deaths, problems of underreporting and multiple reporting are common. Moreover, it is often unreasonable to use the social and economic characteristics of the respondent as a proxy for those of the dead person in order to investigate mortality differentials.

Age misreporting is another serious problem that affects all sources of adult mortality estimates for low- and middle-income countries. Several factors make it difficult to obtain usable information on adult ages and ages at death. Older people are less likely to have birth certificates or health cards than are the young and, in most developing countries, are likely to have received less schooling. Moreover, even if dead persons knew their own age, the informant who reports their death may not. The reported ages of older adults are often exaggerated and ages at death tend to be exaggerated even more. Thus, 'raw' estimates of adult mortality for low- and middle-income countries often require smoothing by fitting a model life table before they can be used to estimate life expectancy or for demographic forecasting, and those on the elderly population may have to be discarded and replaced by data extrapolated from a model life table.

## DATA FOR THE ESTIMATION OF ADULT MORTALITY

A relatively small number – and smaller share by population – of low- and middle-income countries have close to complete registration of adult deaths and population censuses of high quality. A larger number of countries have national or sample vital registration systems that are complete enough to be promising candidates for the methods described in this manual that assess the completeness of registration relative to census counts. In addition, an increasing number of countries have included questions in censuses (or very large sample household surveys) concerning household deaths by

age and sex in some period (most often one year) prior to the census. The completeness of reporting of these deaths can be assessed by the same methods that are used to assess the data on adult deaths collected by registration systems, provided most deaths occur in households and the households don't disintegrate on the death.

A number of countries, particularly in sub-Saharan Africa, have conducted sample surveys (most often under the umbrella of the Demographic and Health Surveys programme) that have included sibling histories that ask each respondent about the survival or otherwise of each of their siblings and when their siblings died. Some countries have sought to measure adult mortality by including questions in censuses and surveys concerning the survival or otherwise of each respondent's mother or father. These data, along with similar summary statistics on siblings, can be tabulated by age of the respondent answering the question and analysed by indirect methods that make use of demographic models to convert them into conventional life table indices of adult mortality.

## DESCRIPTION OF METHODS

Those methods that make use of data on deaths and the population at risk by age (and sex) to estimate adult mortality are collectively referred to as Death Distribution Methods. These methods fall into two distinct groups, depending on how the data are used, the Growth Balance methods and Synthetic Extinct Generations methods (Hill, You and Choi 2009). Both groups of methods require data on deaths from either a registration system or a question in the census together with census-based estimates of the population at risk by age.

The first Growth Balance method is the Brass Growth Balance method developed by Brass (1975), which only requires data on the population by age at a single point in time, but is only applicable if the adult population can be considered to be at least approximately stable (i.e. a population with a regular and unchanging age structure over time). The second Growth Balance method is a generalization of the first method to non-stable populations by Hill (1987), referred to as the Generalized Growth Balance method, which requires data on the population by age at two points in time. The first Synthetic Extinct Generations method is a method developed by Preston, Coale *et al.* (1980), which requires data on the population by age at one point in time and the assumption that the adult population is at least approximately stable. The second Synthetic Extinct Generations method is a generalization of the first approach to non-stable populations by Bennett and Horiuchi (1981; 1984), which requires data on the population at risk at two points in time.

Provided that the assumptions of constant completeness of coverage of the censuses and reporting of deaths by age are reasonably valid, net migration over the period has been small in scale, and there are no major distortions of age reporting between five-year age groups, Death Distribution Methods are the preferred methods for estimating adult mortality both because they provide age-period specific estimates of mortality rates and because they are capable of producing reasonably timely estimates (Hill 2001). However, deciding if these conditions have been met in practice requires a great deal of experience, which means that these methods are amongst the most subjective of the indirect techniques.

One can calculate age-specific death rates directly from counts of deaths and person-years of exposure by age and year derived from sibling histories collected in surveys. As the sample size in surveys such as those conducted by the Demographic and Health Surveys programme is rather small for the estimation of adult mortality, the data should be aggregated into periods of several years. Moreover, as it is common for increasing numbers of dead siblings to be omitted from the histories as the time since their death increases, only mortality estimates for the recent past should be produced from these data.

Data collection instruments such as sibling histories, in which the respondent generally does not live in the same household as the deceased, do not provide a suitable starting point for the collection of data on causes of death. The respondent is unlikely to know the medically-certified cause of death with any precision, especially if the dead person received little or no medical care; alternative approaches, using verbal autopsy methods, which enquire about the signs and symptoms preceding death, will also not work well because the respondent will generally have little first-hand knowledge of such indicators. Some of the same factors apply to information about household deaths collected by censuses, in that household members may well not know the true cause of death, and that the training of interviewers and time available for interviewing each household do not permit detailed probing. However, the census approach can provide a frame for a follow-up verbal autopsy enquiry, on a

sample of households that reported deaths, using carefully-trained interviewers, but such surveys are expensive and complex undertakings.

Two other exceptions to this general rule are that it may be possible to distinguish injury deaths from deaths from natural causes and to identify pregnancy-related deaths, defined as deaths occurring while a woman was pregnant, during childbirth, or during the six weeks after the end of pregnancy. The estimation of pregnancy-related mortality from questions asked both in censuses and during the collection of sibling histories is described in Chapter 29.

The alternative to trying to collect accurate data on deaths and the population at risk by age in order to estimate adult mortality is to use indirect methods of estimation. These methods do not require detailed information on the ages and dates at which people died. Instead, the proportion of individuals remaining alive among some specific category of relative of the respondents answering the question is tabulated according to the age of those respondents. Then, conventional life table measures of survivorship are predicted from these proportions using a regression model fitted to model data in which the relationship between the two quantities is known.

The most successful of the techniques that analyse data on the survival of relatives, estimates the mortality of adult women and men from data on the survival of respondents' mothers and fathers by means of the orphanhood method first developed by Brass and Hill (1973). Contemporary applications of the method usually use the regression coefficients proposed by Timæus (1992) to estimate life table survivorship, rather than the weighting factors proposed initially, as the revised method generates more precise estimates for men. Variants of the method are also discussed here that are intended for use in populations with a high prevalence of HIV infection or when respondents have been asked whether their parents died when the respondent was a child or an adult as indexed, for example, by whether the respondent had married.

If successive sets of data have been collected on maternal or paternal orphanhood in multiple inquiries conducted in the same population, they can be used to estimate adult mortality during the intervening period from synthetic cohort data on orphanhood (Zlotnik and Hill 1981). Such estimates can be made from data on orphans of all ages using the regression coefficients developed for the basic method. This manual, however, focuses on the analysis of synthetic cohort data on orphanhood in adulthood as proposed by Timæus (1991), as this variant of the method is less vulnerable to underreporting of orphanhood by respondents whose natural parent(s) died when the respondent was a young child.

Lastly, the manual describes methods developed by Timæus, Zaba and Ali (2001), that make it possible to estimate adult mortality from data on siblings indirectly if respondents are asked how many of their brothers and sisters survived to adulthood and how many of them have since died.

Other methods for estimating adult mortality indirectly from data on the survival of relatives have been proposed such as asking about the survival of respondents' first husbands and first wives. Experimentation with these questions has shown that respondents often fail to report that they have been widowed. Thus, the method commonly produces severe underestimates of adult mortality. The widowhood method and further methods based on other questions about the survival of relatives that have proved to be unsuccessful are not described in this manual.

A final approach that has been used to estimate adult mortality is the analysis of changes in population size in between two censuses. In a population closed to migration with accurate data, anybody who was present at the first census, but not at the second one, must have died. In practice, except at older ages, the net number of international migrants each year in most countries amounts to a significant fraction of the number of adult deaths. Few countries measure international migration flows accurately enough to adjust for their impact on intercensal population change before estimating mortality. In addition, even small changes in the completeness of the census enumerations can produce severe biases in estimates of adult mortality produced in this way. Thus, in general, the approach cannot be recommended as a method for estimating adult mortality and a detailed account of it is not offered in this manual. Any reader who wishes nevertheless to learn more about this way of estimating mortality is advised to look first at the variant of the approach proposed by Preston and Bennett (1983).

## FURTHER READING AND REFERENCES

No recent paper exists that provides a comprehensive description and assessment of the range of methods available for the estimation of adult mortality in countries with limited and defective data. However, the estimation of adult mortality is discussed alongside child mortality by Hill, Choi and Timæus (2005) and Reniers, Masquelier and Gerland (2011) provide a brief but up-to-date discussion of methods of estimation as well as presenting estimates obtained by putting the methods into practice.


Bennett NG and S Horiuchi. 1981. "Estimating the completeness of death registration in a closed population", *Population Index* **47**(2):207–221.

Bennett NG and S Horiuchi. 1984. "Mortality estimation from registered deaths in less developed countries", *Demography* **21**(2):217–233. doi: 10.2307/2061041

Brass W. 1975. *Methods for Estimating Fertility and Mortality from Limited and Defective Data.* Chapel Hill: International Program of Laboratories for Population Statistics.

Brass W and K Hill. 1973. "Estimating adult mortality from orphanhood," in *International Population Conference, Liège, 1973*. Vol. 3 Liège: International Union for the Scientific Study of Population, pp. 111–123.

Hill K. 1987. "Estimating census and death registration completeness", *Asian and Pacific Population Forum* **1**(3):8–13, 23–24. http://hdl.handle.net/10125/3602

Hill K. 2001. "Methods for measuring adult mortality in developing countries: A comparative review", Paper presented at XXIV IUSSP General Conference, Salvador, Brazil.

Hill K, Y Choi and IM Timæus. 2005. "Unconventional approaches to mortality estimation", *Demographic Research* **13**:281–300. doi: 10.4054/DemRes.2005.13.12

Hill K, D You and Y Choi. 2009. "Death distribution methods for estimating adult mortality: Sensitivity analysis with simulated data error", *Demographic Research* **21**:235–254. doi: 10.4054/DemRes.2009.21.9

Preston SH and NG Bennett. 1983. "A census-based method for estimating adult mortality", *Population Studies* **37**(1):91–104. doi: 10.1080/00324728.1983.10405926

Preston SH, AJ Coale, J Trussell and M Weinstein. 1980. "Estimating the completeness of reporting of adult deaths in populations that are approximately stable", *Population Index* **46**(2):179–202.

Reniers G, B Masquelier and P Gerland. 2011. "Adult mortality trends in Africa," in Rogers, RG and EM Crimmins (eds). *International Handbook of Adult Mortality*. New York, NY: Springer, pp. 151–170. doi: 10.1007/978-90-481-9996-9

Timæus IM. 1991. "Estimation of mortality from orphanhood in adulthood", *Demography* **28**(2):213–227. doi: 10.2307/2061276

Timæus IM. 1992. "Estimation of adult mortality from paternal orphanhood: a reassessment and a new approach", *Population Bulletin of the United Nations* **33**:47–63.

Timæus IM, B Zaba and M Ali. 2001. "Estimation of adult mortality from data on adult siblings," in Zaba, B and J Blacker (eds). *Brass Tacks: Essays in Medical Demography*. London: Athlone, pp. 43–66.

Zlotnik H and K Hill. 1981. "The use of hypothetical cohorts in estimating demographic parameters under conditions of changing fertility and mortality", *Demography* **18**(1):103–122. doi: 10.2307/2061052

# One Census Methods

# Chapter 20 The Brass Growth Balance method

*Rob Dorrington*

## DESCRIPTION OF METHOD

Brass's Growth Balance method (Brass 1975) is the first of what later became known as the Death Distribution Methods for estimating the completeness of the reporting of deaths relative to an estimate of the population. The method makes use of the observation that in a stable population (i.e. a population with an unchanging age structure over time – at least for the adult ages – growing at a constant rate, $r$, each year) that is closed to migration and has accurately reported data, the growth rate, $r$, is equal to the birth rate, $b$, less the death rate, $d$. A similar relationship holds for the population aged $x$ and older, namely, that $r = b(x+) - d(x+)$, where the partial 'birth' rate, $b(x+)$, is defined as the rate at which people turn age $x$ in the population aged $x$ and older and the partial death rate, $d(x+)$, is the rate of mortality of people aged $x$ and older. If, in this population, the deaths are under-reported to the same extent at each age, then $b(x+)\, r + d^r(x+)/c$, where $d^r(x+)$ is the death rate based on the recorded deaths for ages $x$ and older and $c$ is the proportion of deaths that are reported. One can estimate $c$ from the slope of a line fitted to the $b(x+)$, $d^r(x+)$ data points. This estimation is usually confined to adult ages as the (extent of) completeness of reporting of child deaths usually differs from that of adult deaths. Mortality rates can be estimated by dividing the numbers of deaths reported in each age group by $c$ and then dividing these numbers by an estimate of the population exposed to risk based on the population used to estimate the partial birth and death rates.

The method is a particular case of the more general Generalized Growth Balance method, which requires estimates of the population at two points in time but does not require that the population be stable. Readers are referred to that chapter for further detail on the method. It is included in this manual as a method that might be considered when one has an estimate of population numbers only at one point in time.

## DATA REQUIREMENTS AND ASSUMPTIONS

### Tabulations of data required

- Number of deaths of women (men), by five-year age group, and for open age interval $A+$ (with $A$ as high as possible), over a specific period.
- Number of women (men), by five-year age group, and for open age interval $A+$, at or close to the period over which the deaths were measured.

### Important assumptions

- The population is stable, although this assumption can be relaxed to some extent (see below).
- That the completeness of reporting of deaths is the same for all ages above a minimum age (usually age 15).
- The population is closed to migration, although this assumption can be relaxed if net migration is small relative to the mortality rates, or if one has reasonably accurate estimates of the number of migrants by age to allow for in the balance equation (which is very seldom the case).

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Before applying this method, you should investigate the quality of the data at least in the following dimensions:

- age structure of the population;
- sex structure of the population;
- age structure of the deaths; and
- sex structure of the deaths.

## CAVEATS AND WARNINGS

In applying this method, analysts must take particular care with the following:

- The interpretation and estimation processes need to take into account the source (whether vital registration, deaths reported by households in censuses, or deaths recorded at health facilities) of death data as explained

below. However, the biases associated with the source of death data tend to have less impact on the estimate of completeness from the Growth Balance method than on that from the Synthetic Extinct Generations method.
- If applying the method to sub-national geographic areas, the issue of migration typically becomes a greater concern.
- Deciding the age range which is used to fit the straight line to the partial birth and death rates and hence estimate completeness. Issues here are the best age to choose for the open interval if there is evidence of age exaggeration; how to accommodate data points that rise above the line at the older ages because of decreasing completeness with age, possibly due to retirement-associated migration from urban to rural areas where registration is less complete; and whether to exclude ages less than either 30 or 35 because of the impact of migration which has not been allowed for specifically.
- If completeness of reporting of deaths appears to be less than 60 per cent then caution is advised in applying this method as the uncertainty about the estimate is large.

**Application of the method**

***Step 1: Cumulate population and deaths downwards***

To estimate partial birth and death rates one needs to cumulate the numbers in the population and the number of deaths in a defined period of $t$ years for ages $x$ and older. In the case of the population the following equation is used:

$$N(x+) = \sum_{y=x}^{A-5} {}_5N_y + {}_\infty N_A$$

where $A$ is the age at the start of the open age interval.

An analogous equation is used to calculate the number of deaths aged $x$ and older, $D(x+)$.

***Step 2: Calculate the person-years of life lived, PYL(x+)***

In order to estimate partial birth and death rates one needs to estimate the person-years of exposure. This is estimated using the following formula:

$$PYL(x+) = t \cdot N(x+)$$

where $t$ is the length of period over which the deaths have been measured.

***Step 3: Calculate the number of people who turned x in the population, N(x)***

The number of people who turned $x$ (i.e. were 'born' into the open age interval $x+$) in the population is estimated as the geometric mean of the numbers in the two adjacent (five-year) age groups divided by 5, multiplied by the length of the period over which the deaths are reported, in years, using the following formula:

$$N(x) = \frac{t}{5}\left( {}_5N_{x-5} \times {}_5N_x \right)^{\frac{1}{2}}.$$

***Step 4: Calculate partial 'birth' and death rates, b(x+) and d(x+)***

The partial birth and death rates are estimated using the following formulae:

$$b(x+) = \frac{N(x)}{PYL(x+)}$$

and

$$d(x+) = \frac{D(x+)}{PYL(x+)} \text{ respectively.}$$

***Step 5: Plot graph, fit line and estimate completeness, c***

In order to estimate the completeness of reporting of deaths relative to the population, one starts by plotting $b(x+)$ against $d(x+)$ and estimating the coefficients of the straight line fitted to these points, using orthogonal regression, as follows:

$$b = \frac{\sigma_y}{\sigma_x} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \mu_y\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \mu_x\right)^2}$$

and

$$a = \mu_y - b\mu_x.$$

where $b$ is the slope of the line and $a$ the intercept, and $y_i$ represent the $b(x+)$ and the $x_i$ represent the $d(x+)$ and $\mu_y$ and $\mu_x$ represent the means of the two series, respectively.

After fitting the straight line to all the points, one inspects the plotted points relative to the line and the residuals in order to decide on the best range of ages to use to determine the completeness of reporting of deaths. How one decides this is discussed in more detail below, but any points with residuals greater than 1 per cent in absolute value should be excluded. A line is then fitted to the remaining points, and new values of $a$ and $b$ are determined from the fitted line.

The completeness of reporting of deaths, $c$, is derived from these values of $a$ and $b$ as follows:

$$c = \frac{1}{b}\exp\left(a\left(t_c - t_m\right)\right)$$

where $t_c$ is the time of the census and $t_m$ is the mid-point of the period over which the deaths have been recorded. The rationale for this equation is that the reciprocal of the slope estimates the completeness of reporting on the assumption that the census population was at the mid-point of the period over which the deaths have been recorded. In order to correct for any difference between the time of the census and the mid-point of the period over which the deaths have been recorded we need to multiply the estimate of completeness by the ratio of the census population to the estimate of the population at time $t_m$. This is done on the assumption that the population, which is assumed to be stable, is growing at an annual growth rate estimated by $a$. That ratio is $\exp\left(a\left(t_c - t_m\right)\right)$.

***Step 6: Estimate mortality rates adjusted for incompleteness of reporting of deaths***

In order to compute mortality rates one needs first to estimate the population in five-year age groups at the mid-point of the period over which the deaths were recorded by multiplying the census numbers by $\exp\left(-a\left(t_c - t_m\right)\right)$.

Next one needs to adjust the number of deaths for incompleteness by dividing the reported number of deaths by the estimate of completeness, $c$.

The person-years of exposure are estimated by multiplying the estimated population as at $t_m$ by the length of the period over which the deaths were reported, $t$.

Mortality rates adjusted for the incompleteness of the reporting of deaths are thus estimated as follows:

$${}_5m_x = \frac{{}_5D_x / c}{t \times_5 N_x \exp\left(-a\left(t_c - t_m\right)\right)}.$$

Since both the numerator (through the estimate of $c$) and the denominator are adjusted by $\exp\left(-a\left(t_c - t_m\right)\right)$, omitting these adjustments would still produce the same estimates of mortality rates. The estimate of completeness, however, would be equivalent to what it would be if the population at $t_m$ was that at $t_c$.

***Step 7: Smooth using relational logit model life table***

Because the age-specific rates can be erratic they need to be graduated (smoothed). This can be achieved by fitting a Brass relational logit function to a sex-specific standard life table which is considered to have the same shape as that generated by the mortality rates of the population being investigated.

The workbooks (see website) contain a spreadsheet that allows one to produce a smooth set of mortality rates by using a relational logit model fitted to the life table generated by the adjusted mortality rates. The user can choose a standard from the General family of United Nations model life tables or from any of the four families of Princeton model life tables. A custom life table can be entered as standard if there is reason to assume that it better resembles the pattern of adult mortality in the population being studied.

In order to fit the model, probabilities of people aged $x$ dying in the next 5 years, ${}_5q_x$, are estimated from the adjusted rates of mortality as follows:

$${}_5q_x = \frac{5\,{}_5m_x}{1 + 2.5\,{}_5m_x}.$$

From this the life table with a radix of $l_5 = 1$ is calculated as follows:

$$l_{x+5} = l_x\left(1 - {}_5q_x\right).$$

The coefficients, $\alpha$ and $\beta$ are determined by fitting the relational logit model as follows:

$$\gamma_x = \alpha + \beta\gamma_x^s$$

where

$$\gamma_x = 0.5\ln\left(\frac{1 - l_x}{l_x}\right)$$

and superscript '$s$' designates values based on a standard life table.

The fitted life table is then generated from the standard life table using the coefficients $\alpha$ and $\beta$ as follows:

$$\gamma_x^{fitted} = \alpha + \beta\gamma_x^s$$

and

$$l_x^{fitted} = \frac{1}{\exp\left(1 + 2\gamma_x^{fitted}\right)}.$$

The smoothed mortality rates are derived from this life table as follows:

$${}_5m_x^{fitted} = \frac{l_x^{fitted} - l_{x+5}^{fitted}}{T_x - T_{x+5}}$$

and

$$ {}_{\infty}m_x^{fitted} = \frac{l_x^{fitted}}{T_x} $$

where

$$ T_x = \sum_{x=x,5}^{\omega} \frac{5}{2}\left(l_x^{fitted} + l_{x+5}^{fitted}\right) $$

i.e.

$$ T_x = T_{x+5} + \frac{5}{2}\left(l_x^{fitted} + l_{x+5}^{fitted}\right) $$

and $\omega$ is the age above which the life table has no more survivors.

## WORKED EXAMPLE

This example uses data on the numbers of women in the population from the El Salvadorian Census in 1961 and on deaths from vital registration for the calendar year 1961. The example appears in the *BGB_El Salvador* workbook. The reference date for the 1961 Census was midnight between 5 and 6 May, so the date of the census is entered as 1961/05/06 on the *Introduction* sheet.

### *Step 1: Cumulate population, deaths and migrants downwards*

One accumulates the numbers in the population and deaths from the oldest age downwards (Table 20.1).

### *Step 2: Calculate the person-years of life lived, PYL(x+)*

As the deaths are recorded over a single year, the person-years of life lived (column 2 of Table 20.2) are simply the cumulated numbers in the census (i.e. the same as column 4 of Table 20.1) as multiplying by one leaves the numbers unchanged.

### *Step 3: Calculate the number of people who turned x in the population, N(x)*

The numbers of people who turned $x$ are shown in the third column of Table 20.2. For example, the number who turned 70 is estimated as follows:

$$ N(70) = \frac{1}{5}\left(14964 \times 11205\right)^{\frac{1}{2}} = 2589.8. $$

### *Step 4: Calculate partial birth and death rates, b(x+) and d(x+)*

The partial birth and death rates are given in the fourth and fifth columns of Table 20.2. For example these are, for age 20:

$$ b(20+) = \frac{23825}{595352} = 0.0400 $$

and

$$ d(20+) = \frac{5653}{595352} = 0.0095. $$

*TABLE 20.1 Calculation of the cumulated population and deaths, El Salvador, 1961 Census*

| Age | ${}_5N_x$ | ${}_5D_x$ | $N(x+)$ | $D(x+)$ |
|---|---|---|---|---|
| 0–4 | 214,089 | 6,909 | 1,274,253 | 13,652 |
| 5–9 | 190,234 | 610 | 1,060,164 | 6,743 |
| 10–14 | 149,538 | 214 | 869,930 | 6,133 |
| 15–19 | 125,040 | 266 | 720,392 | 5,919 |
| 20–24 | 113,490 | 291 | 595,352 | 5,653 |
| 25–29 | 91,663 | 271 | 481,862 | 5,362 |
| 30–34 | 77,711 | 315 | 390,199 | 5,091 |
| 35–39 | 72,936 | 349 | 312,488 | 4,776 |
| 40–44 | 56,942 | 338 | 239,552 | 4,427 |
| 45–49 | 46,205 | 357 | 182,610 | 4,089 |
| 50–54 | 38,616 | 385 | 136,405 | 3,732 |
| 55–59 | 26,154 | 387 | 97,789 | 3,347 |
| 60–64 | 29,273 | 647 | 71,635 | 2,960 |
| 65–69 | 14,964 | 449 | 42,362 | 2,313 |
| 70–74 | 11,205 | 504 | 27,398 | 1,864 |
| 75+ | 16,193 | 1,360 | 16,193 | 1,360 |

### *Step 5: Plot graph, fit line and estimate completeness, c*

In order to plot the graph and fit the line to all of the data points, one starts by setting the lower age to 5 and the upper age to $A$–1, where $A$ represents the age at the start of the open age interval (75 in this example). The plotted values of $b(x+)$ against $d(x+)$ are shown in Figure 20.1 and the coefficients of the straight line fitted to these points are estimated as follows:

$$ b = \frac{\sigma_y}{\sigma_x} = \frac{0.020547}{0.019103} = 1.0756 $$

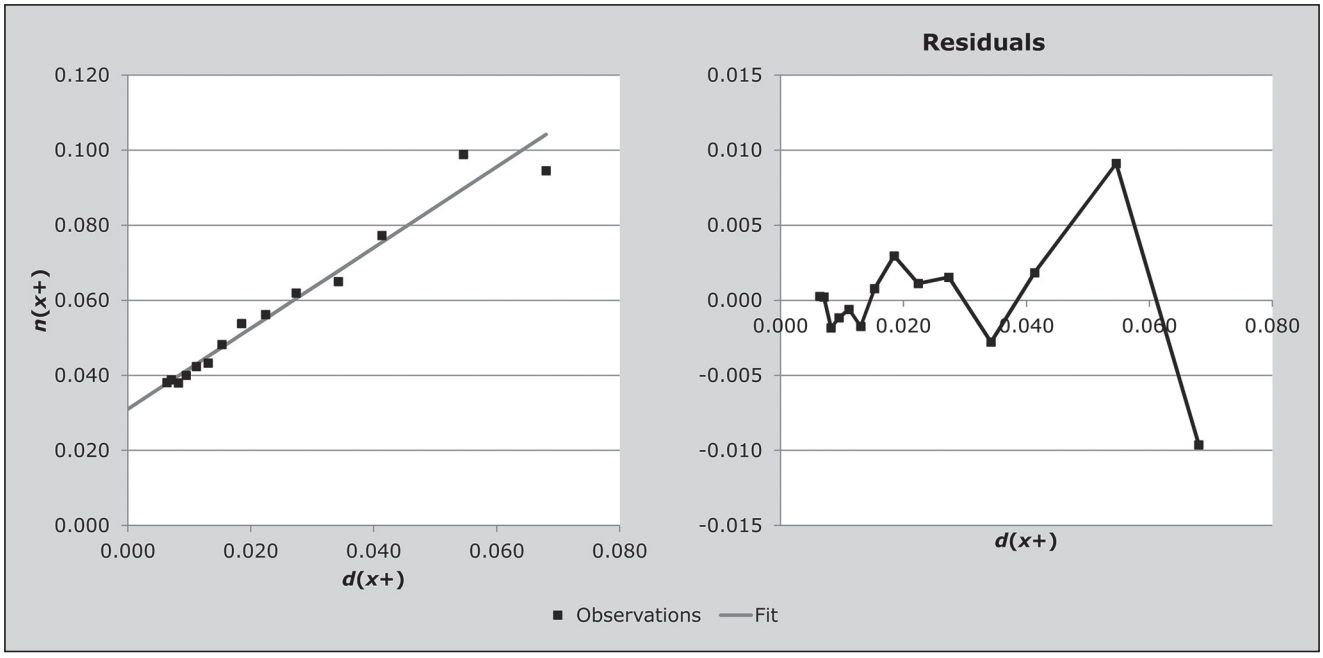
$$ a = 0.05686 - 1.0756 \times 0.02407 = 0.031. $$

Inspection of the diagnostic plots in Figure 20.1 suggests that all but the last (most right-hand) two points lie acceptably close to the fitted line with little evidence of significant migration. Although the residuals of the last two points fall just within the 1 per cent tolerance limits and one could use the estimate of completeness of 93 per cent, it is a useful check to consider the estimate if those two points were dropped.

*TABLE 20.2 Calculation of the person-years lived, the number reaching age x, partial birth and death rates and residuals, El Salvador, 1961 Census*

| Age | $PYL(x+)$ | $N(x)$ | $b(x+)$ | $d(x+) = X$ | $b(x+) = Y$ | $a+bx$ | Residuals $y-(a+bx)$ |
|---|---|---|---|---|---|---|---|
| 0–4 | 1,274,253 | | | 0.00000 | | 0.03097 | |
| 5–9 | 1,060,164 | 40,362 | 0.03807 | 0.00636 | 0.03807 | 0.03782 | 0.00026 |
| 10–14 | 869,930 | 33,733 | 0.03878 | 0.00705 | 0.03878 | 0.03856 | 0.00022 |
| 15–19 | 720,392 | 27,348 | 0.03796 | 0.00822 | 0.03796 | 0.03981 | –0.00185 |
| 20–24 | 595,352 | 23,825 | 0.04002 | 0.00950 | 0.04002 | 0.04119 | –0.00117 |
| 25–29 | 481,862 | 20,399 | 0.04233 | 0.01113 | 0.04233 | 0.04294 | –0.00061 |
| 30–34 | 390,199 | 16,880 | 0.04326 | 0.01305 | 0.04326 | 0.04501 | –0.00175 |
| 35–39 | 312,488 | 15,057 | 0.04818 | 0.01528 | 0.04818 | 0.04741 | 0.00077 |
| 40–44 | 239,552 | 12,889 | 0.05380 | 0.01848 | 0.05380 | 0.05085 | 0.00295 |
| 45–49 | 182,610 | 10,259 | 0.05618 | 0.02239 | 0.05618 | 0.05506 | 0.00112 |
| 50–54 | 136,405 | 8,448 | 0.06193 | 0.02736 | 0.06193 | 0.06040 | 0.00153 |
| 55–59 | 97,789 | 6,356 | 0.06500 | 0.03423 | 0.06500 | 0.06779 | –0.00279 |
| 60–64 | 71,635 | 5,534 | 0.07725 | 0.04132 | 0.07725 | 0.07542 | 0.00183 |
| 65–69 | 42,362 | 4,186 | 0.09881 | 0.05460 | 0.09881 | 0.08970 | 0.00911 |
| 70–74 | 27,398 | 2,590 | 0.09452 | 0.06803 | 0.09452 | 0.10415 | –0.00963 |
| 75+ | 16,193 | | | | | | |

*FIGURE 20.1 Diagnostic plots: open interval of 75+, El Salvador, 1961 Census*



The completeness of reporting of deaths, $c$, is derived from these values of $a$ and $b$ as follows:

$$c = \frac{1}{1.0756}\exp\big(0.031\big(1961.34 - 1961.5\big)\big)$$
$$= \frac{1}{1.0756}e^{0.031\times 0.16} = 92.51\%.$$

Dropping the last two points (by setting the upper age of the chosen range to 64) produces the diagnostic plots shown in Figure 20.2 and an estimate of completeness of 89 per cent, which is sufficiently close to suggest that it is unnecessary to drop the last two points. Dropping only the last point produces a big change in the estimate (to 82 per cent) and a poorer fit for some of the points to the left, which suggests that this is probably not the best course of action. As a general rule, it is not recommended in a population with significant digital preference to truncate at an age ending in zero.

***Step 6: Estimate mortality rates adjusted for incompleteness of reporting of deaths***

The population as at the mid-point of the period over which the deaths were recorded is estimated by adjusting the census population for the growth between the two dates at the estimated growth rate of 3.1 per cent. These estimates are shown in the second column of Table 20.3. For example for the 15–19 age group the number is estimated as follows:

$$_5N_{15}(t_m) = 125040 \times \exp\big(0.031\big(1961.34 - 1961.50\big)\big)$$
$$= 125040e^{0.031\times 0.16} = 125662.3.$$

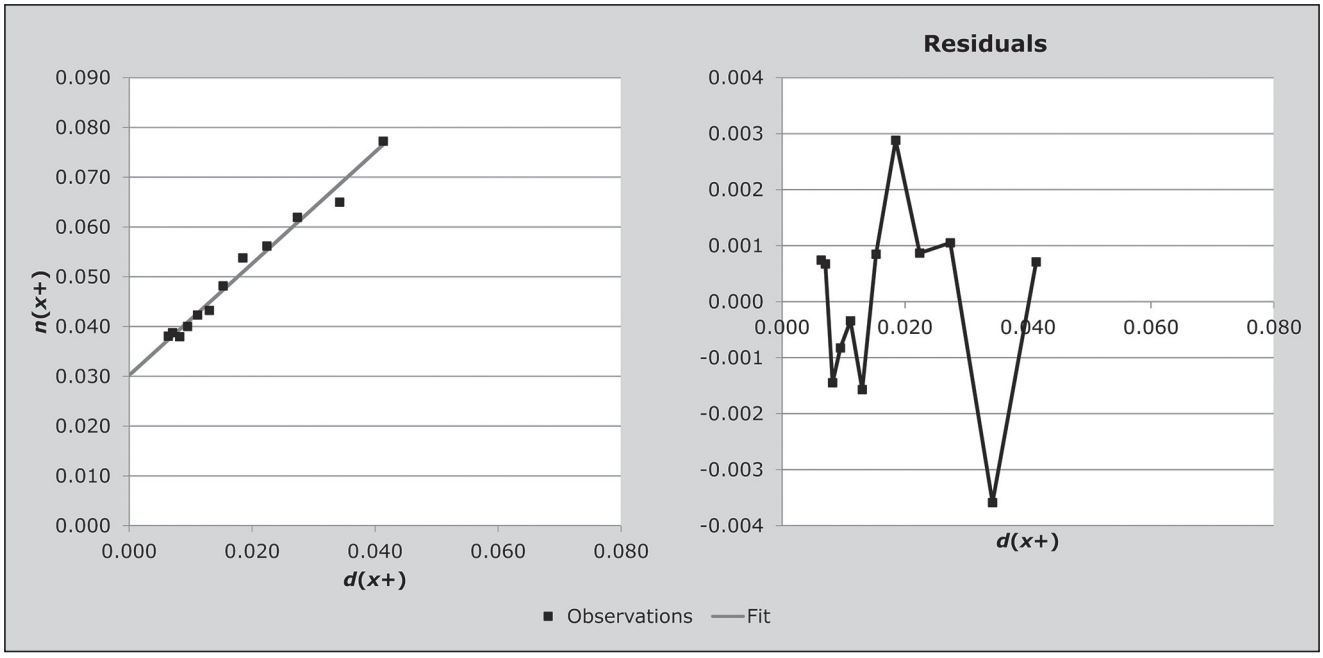
Next the deaths are adjusted for incompleteness by dividing the number of reported deaths in each age group by the estimate of completeness. These numbers are shown in column 3 of Table 20.3. For example, for the 15–19 age group the number is derived from the number of reported deaths (shown in column 3 of Table 20.1), 266, as follows:

$$\frac{266}{0.9251} = 287.5\,.$$

The adjusted person-years of life lived (column 4 of Table 20.3) are the numbers in the population at the mid-point of the period over which the deaths have been recorded (column 2 Table 20.3) multiplied by the length (in years) of the period over which the deaths are recorded, which in this case is 1 year.

The mortality rates adjusted for incompleteness of

*FIGURE 20.2 Diagnostic plots: open interval of 65+, El Salvador, 1961 Census*



reporting of deaths (column 5 of Table 20.3) are derived by dividing the adjusted deaths by the adjusted person-years of life lived. For example, for the 15–19 age group the adjusted rate is calculated as follows:

$$\frac{287.60}{125662} = 0.0023.$$

*TABLE 20.3 Calculation of adjusted mortality rates, El Salvador, 1961 Census*

| Age | Adjusted ${}_5N_x(t_m)$ | Adjusted ${}_5D_x$ | Adjusted $PYL(x,5)$ | Adjusted ${}_5m_x$ |
|---|---|---|---|---|
| 0–4 | | | | |
| 5–9 | 191,181 | 659 | 191,181 | 0.0034 |
| 10–14 | 150,282 | 231 | 150,282 | 0.0015 |
| 15–19 | 125,662 | 288 | 125,662 | 0.0023 |
| 20–24 | 114,055 | 315 | 114,055 | 0.0028 |
| 25–29 | 92,119 | 293 | 92,119 | 0.0032 |
| 30–34 | 78,098 | 340 | 78,098 | 0.0044 |
| 35–39 | 73,299 | 377 | 73,299 | 0.0051 |
| 40–44 | 57,225 | 365 | 57,225 | 0.0064 |
| 45–49 | 46,435 | 386 | 46,435 | 0.0083 |
| 50–54 | 38,808 | 416 | 38,808 | 0.0107 |
| 55–59 | 26,284 | 418 | 26,284 | 0.0159 |
| 60–64 | 29,419 | 699 | 29,419 | 0.0238 |
| 65–69 | 15,038 | 485 | 15,038 | 0.0323 |
| 70–74 | 11,261 | 545 | 11,261 | 0.0484 |
| 75+ | 16,274 | 1,470 | 16,274 | 0.0903 |

***Step 7: Smooth using relational logit model life table***

Estimates of probabilities of women aged $x$ dying in the next 5 years, ${}_5q_x$, estimated from the adjusted rates of mortality, are shown in the second column of Table 20.4. For example, the probability of a 15-year old woman dying before reaching age 20 is calculated as follows:

$${}_5q_{15} = \frac{5\times 0.00229}{1+2.5\times 0.00229} = 0.0114.$$

The life table proportions of five-year olds alive at age $x+5$ estimated from the proportion alive at age $x$ using these values appear in column 3 of Table 20.4. For example, the proportion alive at age 20 is calculated as follows:

$$l_{20} = 0.9754(1-0.0114) = 0.9643.$$

The logit transformations of the proportions surviving appear in column 4 of Table 20.4. For example, the logit transformation of the $l_{20}$ is calculated as follows:

$$\gamma_{20} = 0.5\ln\left(\frac{1-0.9643}{0.9643}\right) = -1.6477.$$

The logit transformation of the conditional life table for females based on the West family of Princeton model life tables with $e0$=60 in column 5 of Table 20.4 appears in column 6 of Table 20.4. As can be seen from Figure 20.3, the West model appears to fit the data well, with the possible exception of the youngest ages.

The coefficients $\alpha$ and $\beta$ are determined as the intercept and slope, respectively, of the straight line fitted to the logit transformations in columns 4 and 6 of Table 20.4 over the range of ages chosen by the user (45 and 75 in this example), namely 0.0211 and 0.9672 respectively.

These coefficients are then applied to the logit transformation of the conditional model life table to produce the fitted logits in column 7 of Table 20.4. Thus, for example, the fitted logit at age 20 is calculated as follows:

$$\gamma_{20}^{fitted} = -0.0211 + 0.9672\times(-1.7060) = -1.671.$$

These values are then used to produce the fitted life table in column 8 of Table 20.4. For example, the value at age 20 is calculated as follows:

$$l_{20}^{fitted} = \frac{1}{\exp(1+2\times(-1.671))} = 0.9658.$$

The conditional years of life lived, $T_x$, which appear in column 9 of Table 20.4 are then calculated from the fitted life table and these numbers are used to produce the smoothed mortality rates which appear in column 10 of Table 2.4. For example, for age 80:

$$T_{80} = 0.577 + \frac{5}{2}(0.2695+0.132) = 1.575$$

$${}_5m_{80}^{fitted} = \frac{.2695-0.132}{1.575-0.571} = 0.137.$$

## DIAGNOSTICS, ANALYSIS AND INTERPRETATION

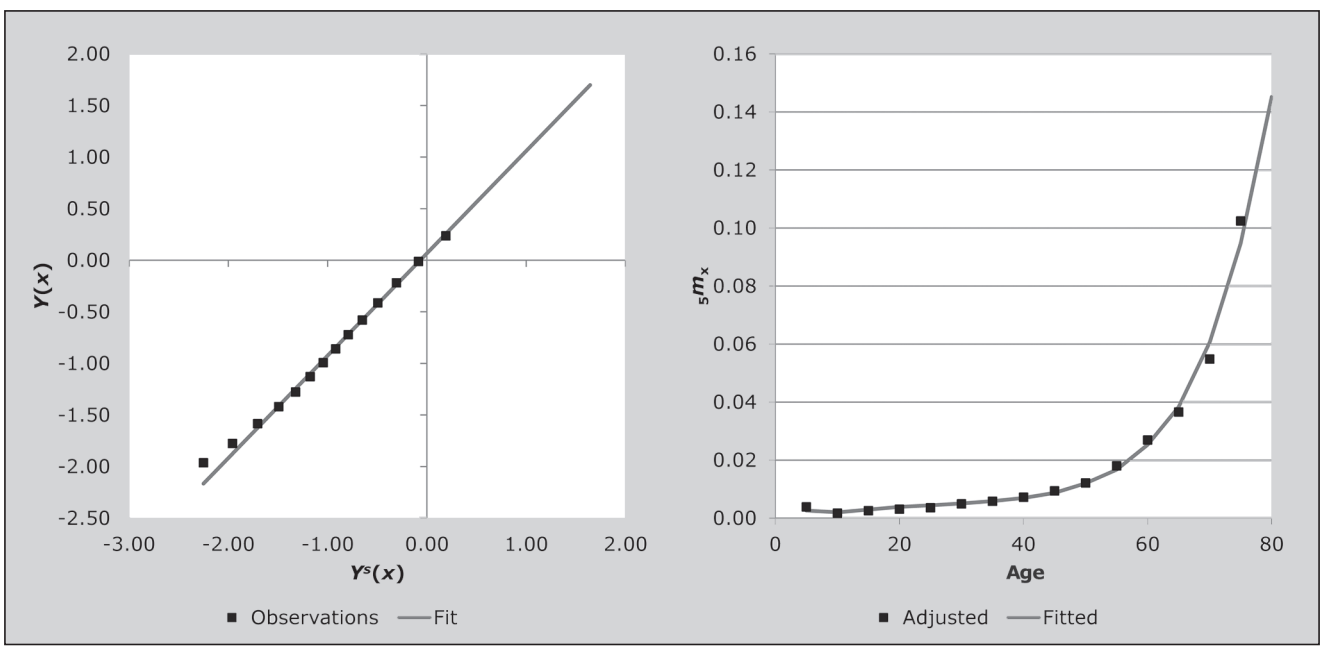
### Checks and validation

The example above was taken from *Manual X* (UN Population Division 1983) which produced an estimate of completeness of around 83 per cent from application of both this method

*TABLE 20.4 Calculation of smoothed mortality rates using a relational logit model life table, El Salvador, 1961 Census*

| Age | ${}_5q_x$ | $l_x/l_5$ | Obs. $Y(x)$ | Princeton West $l^s(x)$ | $Y^s(x)$ | Fitted $Y(x)$ | Fitted $l(x)$ | $T(x)$ | Smooth ${}_5m_x$ |
|---|---|---|---|---|---|---|---|---|---|
| 0 | | | | | | | | | |
| 5 | 0.0171 | 1 | | 1.0000 | | | 1 | 61.957 | 0.0025 |
| 10 | 0.0077 | 0.9829 | –2.0258 | 0.9890 | –2.2506 | –2.1978 | 0.9878 | 56.987 | 0.0018 |
| 15 | 0.0114 | 0.9754 | –1.8394 | 0.9805 | –1.9585 | –1.9153 | 0.9788 | 52.071 | 0.0027 |
| 20 | 0.0137 | 0.9643 | –1.6477 | 0.9681 | –1.7060 | –1.6710 | 0.9658 | 47.209 | 0.0035 |
| 25 | 0.0158 | 0.9511 | –1.4836 | 0.9519 | –1.4928 | –1.4649 | 0.9493 | 42.421 | 0.0039 |
| 30 | 0.0216 | 0.9361 | –1.3419 | 0.9337 | –1.3226 | –1.3003 | 0.9309 | 37.721 | 0.0045 |
| 35 | 0.0254 | 0.9159 | –1.1938 | 0.9132 | –1.1766 | –1.1590 | 0.9104 | 33.118 | 0.0051 |
| 40 | 0.0314 | 0.8926 | –1.0588 | 0.8899 | –1.0447 | –1.0314 | 0.8872 | 28.624 | 0.0061 |
| 45 | 0.0407 | 0.8646 | –0.9269 | 0.8628 | –0.9194 | –0.9103 | 0.8606 | 24.254 | 0.0076 |
| 50 | 0.0522 | 0.8294 | –0.7906 | 0.8299 | –0.7925 | –0.7875 | 0.8285 | 20.031 | 0.0105 |
| 55 | 0.0765 | 0.7861 | –0.6507 | 0.7863 | –0.6514 | –0.6511 | 0.7862 | 15.994 | 0.0146 |
| 60 | 0.1122 | 0.7259 | –0.4870 | 0.7289 | –0.4946 | –0.4995 | 0.7308 | 12.202 | 0.0222 |
| 65 | 0.1493 | 0.6445 | –0.2974 | 0.6490 | –0.3074 | –0.3184 | 0.6540 | 8.740 | 0.0339 |
| 70 | 0.2158 | 0.5482 | –0.0968 | 0.5427 | –0.0856 | –0.1039 | 0.5517 | 5.725 | 0.0545 |
| 75 | #N/A | 0.4299 | 0.1411 | 0.4062 | 0.1898 | 0.1625 | 0.4194 | 3.297 | 0.0871 |
| 80 | #N/A | #N/A | #N/A | 0.2545 | 0.5373 | 0.4986 | 0.2695 | 1.575 | 0.1370 |
| 85 | #N/A | #N/A | #N/A | 0.1201 | 0.9956 | 0.9419 | 0.1320 | 0.571 | 0.2084 |

*FIGURE 20.3 Observed logits and adjusted mortality rates against expected derived from the female West model life table, El Salvador, 1961 Census*



and the Preston and Coale method. The difference between these estimates and the one produced in this application (93 per cent) appears to be due largely to the method used to fit the line combined with the points used to fit the line. The *Method* sheet in the *BGB_El Salvador* workbook (see website) uses orthogonal regression while *Manual X* applied 'grouped means' to points up to age 60, and 'trimmed means' – thus effectively removing the impact of the final data point. This difference suggests that a case could be made for dropping the last two points in the example on the grounds that the regression is unduly influenced by points at the extremes of the axes. However, as indicated above, the effect of doing this when using orthogonal regression to fit the line is not particularly significant.

**Interpretation**

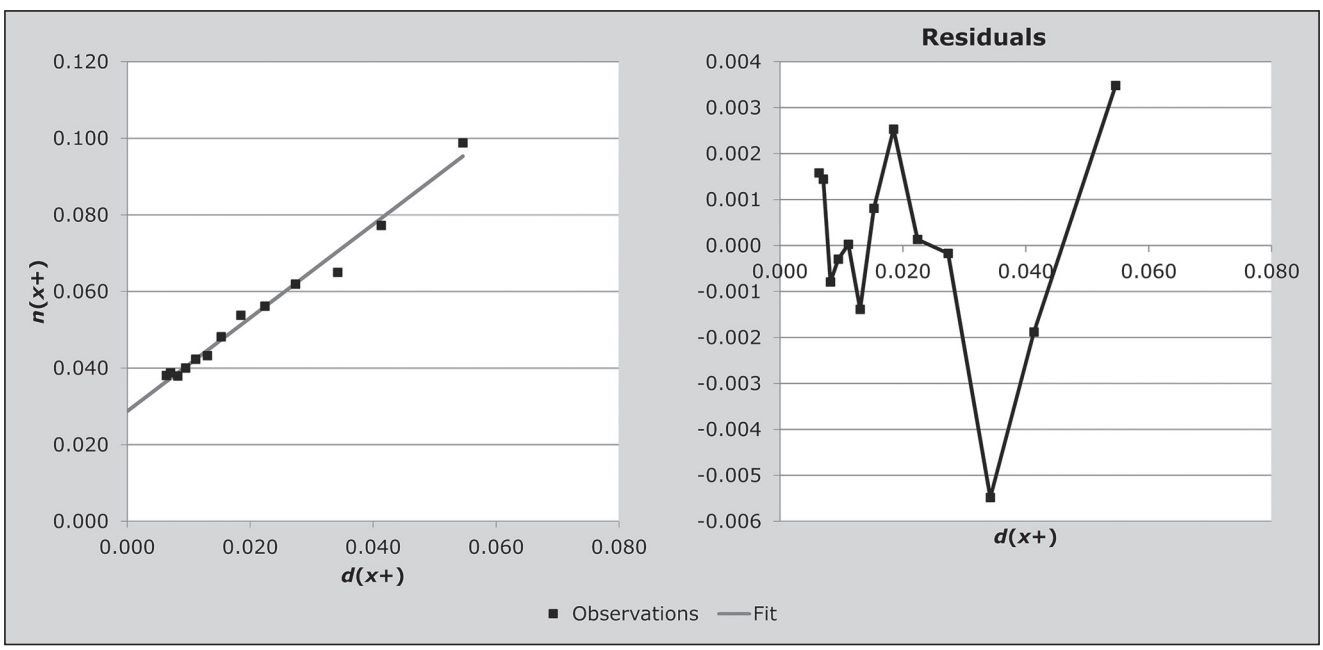
A problem that often arises with deciding on the 'upper age' for fitting the straight line is that estimates of completeness may vary quite considerably due to the exclusion of a single point. For example if one were to choose 70+ as the open interval the diagnostic plots would look as shown in Figure 20.4 and the estimate of completeness would be 82 per cent. The diagnostic plots, on their own, do not suggest that this fit is particularly worse than that using a 65+ open interval. In such cases one should calculate the estimate for several open intervals and use one which represents the estimate of completeness closest to the majority, or the median. Thus in this case the estimate of completeness for the open interval 60+ is 91 per cent, suggesting that the deaths are around 90 per cent complete. However, as pointed out above, as a general rule, it is not recommended in a population with significant digital preference to truncate at an age ending in zero.

## METHOD-SPECIFIC ISSUES WITH INTERPRETATION

**Source of reported deaths**

Generally there are two sorts of problems with the deaths data: those that lead to under/over coverage that is constant by age, which is precisely what the method is intended to address, and those which lead to differential coverage by age, which can distort the estimates. Although the general approach remains essentially the same irrespective of the source of the death data, different sources of data are prone to different biases which might impact on the interpretation of the results. These differences are illustrated by way of particular examples, but, in general terms, you need to look out for the following biases in the death data.

***FIGURE 20.4*** *Diagnostic plots: open interval of 70+, El Salvador, 1961 Census*



***1) Vital registration***

If the proportionate split of the population between urban and rural (or appropriate proxies) areas differs significantly by age and the completeness of reporting of deaths in urban areas is significantly higher than it is in rural areas, then the assumption that completeness is independent of age is likely to be violated by a falling off of completeness with age at ages over 50, if a proportion of people move from urban to rural areas on retirement. If ignored, this violation is likely to lead to an underestimate of the average level of completeness.

***2) Deaths reported by households in censuses/surveys***

The data are subject to three potential problems:

- If a significant proportion of households dissolve on the death of a key person (e.g. the sole breadwinner), then the deaths of such people go unreported, leading to a violation of the assumption that completeness is invariant with age. If a significant proportion of deaths in some age groups are of individuals who do not live in private households (for example, they live in homes for the elderly), the breach of the assumption could be even more severe. However, this is not an issue in most developing countries.
- In situations where young adults leave the home they grew up in to work in urban areas, it is possible that they are regarded as being members of more than one household (or of neither household) and their deaths could be reported more than once (or not at all), again leading to a violation of the assumption of constant reporting of deaths by age. In this case, one can limit the impact by ignoring the data below a specific age in determining completeness.
- Reference period error: Since there is often confusion about the exact period for which deaths are to be reported, in addition to uncertainty about exact dates of death, it is possible for there to be overall under- or over-reporting of deaths. Provided one can assume that this is independent of the age of the deceased, this distortion will be accounted for in the estimate of completeness and is not a problem for estimating mortality rates.

***3) Deaths recorded in health facilities***

Little is known about how well this source of data works. However, it can be expected that completeness would depend on the distribution of health services from which the data have been gathered, and in many developing countries such services are likely to be concentrated in urban areas. So again, if the proportion of the population living in urban rather than rural areas varies with age, then completeness cannot be assumed to be independent of age. It is also possible that certain causes will predominate in facilities and, if these causes are significant and age-related, this could lead to a further violation of the assumption of constant completeness by age.

In all such cases, the plotted points will lie progressively above the fitted line at the older ages leading to an underestimate of completeness. The estimate will be improved, although still biased downward slightly, by excluding the points at the highest ages from determining the fitted line.

## DETAILED DESCRIPTION OF METHOD

### Mathematical exposition

Although the Brass Growth Balance method is simply a special case of the Generalized Growth Balance method, with the growth rate of the population aged $x+$, $r(x+)$ constant for all ages, it might be of assistance to understanding these methods to describe the specific case as well.

Brass's Growth Balance method (Brass 1975) has its origins in work by Carrier (1958) who first proposed a way of estimating mortality from the age distribution of deaths. The method derives from the relationship found in the balancing equation for a population closed to migration. In such a population the number of people in the population at time $t_2$ is the number at time $t_1$, plus the births that have occurred between time $t_1$ and $t_2$, less the deaths that have occurred between times $t_1$ and $t_2$, i.e. ${}_{\infty}N_0(t_2) = {}_{\infty}N_0(t_1) + B - D$, where $B$ and $D$ are the births and deaths, respectively, that occurred between times $t_1$ and $t_2$. This equation can be generalized to hold for any population aged $x$ and older, provided we have an estimate of the number of people who turned $x$ (i.e. joined the age interval through aging) between the times $t_1$ and $t_2$, $N_x$, and the number of deaths aged $x$ and older that occurred between times $t_1$ and $t_2$, ${}_{\infty}D_x$. Thus

$$ {}_{\infty}N_x(t_2) = {}_{\infty}N_x(t_1) + N_x - {}_{\infty}D_x. \qquad (1) $$

If we rewrite Equation 1 as

$$ {}_{\infty}N_x(t_2) - {}_{\infty}N_x(t_1) = N_x - {}_{\infty}D_x $$

and divide through by the person-years of exposure between times $t_1$ and $t_2$ we can express the balance equation in terms of rates, i.e.

$$r(x+) = b(x+) - d(x+), \quad (2)$$

where

$$r(x+) = \frac{{}_{\infty}N_x(t_2) - {}_{\infty}N_x(t_1)}{\int_{t_1}^{t_2} {}_{\infty}N_x(t)dt},$$

$$b(x+) \approx \frac{N_x}{\int_{t_1}^{t_2} {}_{\infty}N_x(t)dt}$$

and

$$d(x+) \approx \frac{{}_{\infty}D_x}{\int_{t_1}^{t_2} {}_{\infty}N_x(t)dt}.$$

$b(x+)$ and $d(x+)$ are often referred to as *partial* or *segmental* birth and death rates, respectively.

These relationships only hold if there is complete and accurate recording of birthdays and deaths by age between times $t_1$ and $t_2$, and counting of the population by age at times $t_1$ and $t_2$.

Now suppose that all data are accurate except that the deaths are incompletely reported. Suppose further that one can assume (at least above a certain age – typically confined to adult ages) that a fixed proportion, $c$, of deaths are reported independent of the age of the deceased. Then ${}_{\infty}D_x = {}_{\infty}D_x^r / c$, where ${}_{\infty}D_x^r$ represents the number of reported deaths aged $x$ and older, and Equation 2 becomes $r(x+) = b(x+) - d^r(x+)/c$, where

$$d^r(x+) = \frac{{}_{\infty}D_x^r}{\int_{t_1}^{t_2} {}_{\infty}N_x(t)dt}.$$

If we further assume that the population is stable, growing at a constant rate of $r$ a year, then this equation can be rearranged as follows: $b(x+) = r + kd^r(x+)$. Thus, if one fits a straight line to the points $(b(x+), d^r(x+))$, the intercept provides an estimate of the growth rate, $r$, and the reciprocal of the slope, $k$, provides an estimate of the completeness of reporting of the deaths, $c$.

Mortality rates by age group, ${}_5m_x$, are then estimated as

$${}_5m_x = \frac{{}_5D_x^r / c}{\int_{t_1}^{t_2} {}_{\infty}N_x(t)dt}.$$

**Implementation of the method**

Assume that in practice one has data on the following: the number of reported deaths over a number of years, from times $t_1$ to $t_2$, in five-year age groups, ${}_5D_x^r$, up to an open interval at age $A$, ${}_{\infty}D_A^r$; and the number of people in the population in the middle of this period, in the same age groups, ${}_5N_x$ up to ${}_{\infty}N_A$. These data can then be used to apply the method by computing ${}_{\infty}D_x^r$ and ${}_{\infty}N_x$, and approximating $N_x = \left({}_5N_{x-5} + {}_5N_x\right)/10$ or $N_x = \left({}_5N_{x-5} \cdot {}_5N_x\right)^{\frac{1}{2}}/5$ and $\int_{t_1}^{t_2} {}_{\infty}N_x(t)dt = {}_{\infty}N_x\left(t_2 - t_1\right)$.

If, instead of the population in the middle of the period, one had the population at some other time, say $t$, then one can apply the method using that population instead. The only difference is that the estimate of completeness will be relative to the population at time $t$ as if it was the population as at the midpoint of the period. In other words, assuming the population to be stable, ${}_{\infty}N_x(t) = {}_{\infty}N_x e^{r\left(t-(t_2-t_1)/2\right)}$ and the completeness relative to this population is $ce^{r\left(t-(t_2-t_1)/2\right)}$, the mortality rates derived by dividing the reported number of deaths corrected for this level of completeness by ${}_{\infty}N_x(t)$ will give the same rates as one would obtain if one had had estimates of the population by age group in the middle of the period.

**Fitting of the straight line**

There are two aspects to determining the straight line that best represents the relationship between the partial birth and death rates, namely, the choice of method and the choice of points used to determine the slope and intercept.

Fitting the straight line using unweighted least squares regression is not recommended since this method gives too much weight to the values at the older ages, which tend to be less reliable. Thus, it is recommended that one fit the line using a more robust method such as the 'mean' line (i.e. the line defined as that joining the two points represented by the mean of the vertical axis values and the mean of the horizontal axis values of the first half and the second half of the age range) or the 'trimmed mean' line (i.e. the same as the mean line except that the average of the points is a weighted average – weighting the less reliable points, usually at the extremes, less than the other points). These methods are explained in detail in the *Manual X* (UN Population Division 1983: 144–145). An alternative is described in more detail in the *UN Manual on Adult Mortality* (UN Population Division

2002: 105–110). The alternative is similar to the 'mean' line, except that one splits the range of points into three equally sized groups,[1] and determines the line that joins the medians of the independent and dependent variables in the lowest third and the highest third of points.

Bhat (2002) points out that each method has its drawbacks and suggests, since it does not matter whether the partial birth or partial death rates are treated as dependent variable, that orthogonal regression is the best method for dealing with age misstatement. This reflects both vertical and horizontal distance from the line (by minimizing the orthogonal residual sum of squares (ORSS) $\sum_i \frac{x_i^2 y_i^2}{x_i^2 + y_i^2}$).

Using this method *c*, the completeness of the death reporting, is estimated as the ratio of the standard deviation of the partial death rates to the standard deviation of the partial birth rates. The intercept is the mean of the partial birth rates, minus the mean of the partial death rates divided by *c*. This is the approach used in the applications of the Brass Growth Balance method in the workbooks (see website).

**Limitations**

The major limitations of the method as described above and provided for in the spreadsheet are that it requires that the population be stable and closed to migration and it should not be applied when these conditions do not apply to any significant extent. By way of example of inappropriate usage, application of this method (data available in the *GGB_South Africa_males* workbook) to estimate completeness of reporting of deaths in South Africa between the 2001 Census and a census replacement survey in 2007, estimating the population in the middle of the period as the average of the two survey populations, provides an estimate of completeness, using the same age range, of 85 per cent. Increasing the minimum age of range of the data used to fit the straight line to 35 increases the estimate to 88 per cent, still somewhat lower than the estimate of 92 per cent produced using the Generalized Growth Balance method.

This method is less vulnerable to age misreporting than the Preston and Coale method. In particular, for example, the common tendency to exaggerate the age reported at death (relative to that recorded at census) will manifest itself by the plotted points falling off to the right (i.e. below the fitted line) over the range of exaggerated ages and this can be allowed for when deciding which points to use to fit the line. The method is, however, more vulnerable to the effects of destabilization resulting from a rapid decline in mortality (Martin 1980), in which case it tends to underestimate the extent of completeness since the lighter mortality is "interpreted" by the model as increased under-reporting (i.e. steeper slope). However, simulation has shown (Rashad 1978) that the bias resulting from a slow steady improvement in mortality (as has been experienced by some developing countries in the absence of epidemics, famine and wars) is quite small.

As far as changes in fertility rates are concerned, provided these have occurred not more than 15 years ago these changes will have little impact on the performance of the method since they affect mainly the youngest age groups.

Migration is likely to affect the young adult population (mainly between 20 and 35) but to have much less effect on deaths, which largely occur in old age. Unaccounted-for immigration will tend to lower the slope and hence lead to an over-estimate of the extent of death registration and an underestimate of mortality rates. Unaccounted-for emigration will have the opposite effect. Some demographers advocate fitting the straight line to data down to age 5 to limit the effect of unaccounted-for migration, on the assumption that any differences in completeness of reporting of deaths at these younger ages from that of the older ages is unlikely to lead to any major distortions since mortality is very light between ages 5 and 14. However, it is doubtful that this adaptation removes much of the bias.

Alternatively one could confine the fit to points above age 35 to remove the bulk of the effect of migration. However, often the data at the older ages is more suspect making the estimate of completeness less reliable. Although using these adaptations probably produces better estimates than simply ignoring migration, there is, unfortunately, little research into the accuracy of the estimated completeness produced by these adaptations.

Technically, if one had reliable estimates of net migration by age, one could adapt the method by replacing the partial birth rate $b(x+)$ by $b(x+) - i(x+)$, where $i(x+)$ is the net in-migration rate, in deriving the fitted line. However, in practice, in situations where one has to apply this method one rarely has sufficiently reliable estimates of net migration by age to warrant adapting the method.

1 Where it is not possible to divide the total number of points, *n*, into three equally sized groups then the highest and lowest group are taken as the top and bottom, respectively, int(*n*/3)+1 points.

Fluctuations in the completeness of death registration with age are likely to introduce curvature in the pattern of points. Consequently, one of the strengths of this method is that if the points for successive age boundaries fall on a reasonably straight line, then it is probably reasonable to assume that completeness is constant with respect to age. However, where some but not all the points lie on a straight line one way of deciding which points to discard is to calculate the segmental growth rate for each successive open interval and then use those points for which the values of $r_{a+}$ are reasonably consistent.

Perhaps the most important limitation of the method is that the plot of partial birth rates against partial death rates is, with the exceptions mentioned above, diagnostically quite limited. In particular, simulations of a demographically stable population which then suffers an increase in mortality due to HIV/AIDS with a population prevalence of 11 per cent, produce a plot of points which to all intents and purposes fit a straight line but underestimate the level of completeness, even if one confines the fit to ages over 45. The lesson is that, if the points do not fall on a straight line, there are problems with the data; however, if the points do fall on straight line, you cannot be certain that the estimates of completeness and coverage are correct.

**Extensions**

If one had accurate data and a reliable, independent, estimate of $r$ (and the population is stable) then one could reformulate Equation 2 to estimate $c(x+)$, the completeness of each open-ended age group, as follows: $c(x+) = {}_{\infty}D_x^r \Big/ \left( N_x - r\,{}_{\infty}N_x \right)$. However, in practice it is rare to find a sufficiently stable population with sufficiently accurate age reporting to make such an exercise worthwhile.

## FURTHER READING AND REFERENCES

Since this method is a particular case of the more general Generalized Growth Balance method, readers are referred to Chapter 24 for further reading.


Bhat M. 2002. "General Growth Balance method: A reformulation for populations open to migration", *Population Studies* **56**(1):23–34. doi: 10.1080/00324720213798

Brass W. 1975. *Methods for Estimating Fertility and Mortality from Limited and Defective Data.* Chapel Hill NC: Carolina Population Centre.

Carrier NH. 1958. "A note on the estimation of mortality and other population characteristics, given death by age", *Population Studies* **12**:149–163. doi: 10.2307/2172187

Martin LG. 1980. "A modification for use in destabilized populations of Brass's Technique for estimating completeness of death registration", *Population Studies* **34**:381–395. doi: 10.2307/2175194

Rashad HM. 1978. "The Estimation of Adult Mortality from Defective Registration Data." Unpublished PhD thesis, London: University of London.

UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

UN Population Division. 2002. *Methods for Estimating Adult Mortality*. New York: United Nations, Department of Economic and Social Affairs, ESA/P/WP.175. http://www.un.org/esa/population/techcoop/DemEst/methods_adultmort/methods_adultmort.html

# Chapter 21 The Preston and Coale method

*Rob Dorrington*

## DESCRIPTION OF METHOD

The Preston and Coale method (Preston, Coale, Trussell *et al.* 1980) is the second of what later became known as the Death Distribution Methods for estimating the completeness of the reporting of deaths relative to an estimate of the population at one point in time. It makes use of the observation that the number of people of a given age alive at a point in time must be equal to the number of people from that cohort who die from that point in time onward. If the population is stable (i.e. a population with an unchanging age distribution – at least for adult ages – growing at a constant rate, $r$, each year) and closed to migration, and the reported data are accurate, the number of deaths aged $x$, $t$ years in the future, will equal the number of deaths aged $x$ currently, multiplied by $e^{rt}$. It is thus possible to estimate the current population aged $y$ using only current deaths by age above age $y$ and the stable growth rate $r$. If the number of current deaths is under-reported, but can be assumed to be under-reported to the same extent, $c$, at every age, then the estimate of the future number of cohort deaths will be underestimated to the same extent. Thus, it is possible to estimate the completeness of reporting of deaths by dividing the sum of the estimates of future cohort deaths derived from the number of deaths at any date by the population at the same date. Mortality rates can then be estimated by dividing the numbers of deaths reported in each adult age group by $c$ and then dividing these numbers by an estimate of the population exposed to risk.

The method is a particular case of the more general Synthetic Extinct Generations method, which requires estimates of the population at two points in time but does not require that the population is stable. Readers are referred to that chapter for further detail on the method. It is included in this manual as a method that might be considered when one has an estimate of population numbers at only one point in time.

## DATA REQUIREMENTS AND ASSUMPTIONS

### Tabulations of data required

- Number of deaths of women (men), by five-year age group, and for open age interval $A$+ (with $A$ as high as possible), over a specific period.
- Number of women (men), by five-year age group, and for open age interval $A$+, at or close to the period over which the deaths were measured.

### Important assumptions

- The population is stable, although this assumption can be relaxed to some extent (see below).
- The completeness of reporting of deaths is the same for all ages above a minimum age (usually age 15).
- The population is closed to migration, although this assumption can be relaxed if net migration is small relative the mortality rates, or if one has reasonably accurate estimates of the number of migrants by age to allow for in the balance equation (which is very seldom the case).

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Before applying this method, you should investigate the quality of the data at least in the following dimensions:

- age structure of the population;
- sex structure of the population;
- age structure of the deaths; and
- sex structure of the deaths.

## CAVEATS AND WARNINGS

In applying this method, analysts must take particular care with the following.

- The interpretation and estimation processes need to take into account the source of death data (vital registration, reported by households in censuses, or deaths in hospitals) as explained below.

- If applying the method to sub-national geographic areas, the issue of migration typically becomes a greater concern.
- Deciding the age range which is to be used to determine the growth rate (i.e. such that the age-specific estimates of completeness minimize the absolute difference from the mean estimate of completeness). Issues here are whether the best estimate of the growth rate to use is the intercept determined as a result of applying the Brass Growth Balance method to the same data (which would be the case if completeness was thought to decrease among the elderly, perhaps associated with retirement), and whether to exclude ages below 30 or 35 because of the impact of migration which has not been allowed for specifically.
- Deciding on the age range to use for determining the estimate of completeness. Typically, this range might exclude young adults if there is significant unaccounted-for migration, and old people if the results suggest that fewer of their deaths are reported than deaths of younger adults.
- Ensuring that the *Solver* routine in Excel has run satisfactorily (i.e. has produced a sensible result). Occasionally *Solver* offers a solution which is manifestly too low. In such situations it is best to adjust delta manually in the right direction and apply *Solver* to this new starting value.
- Ensuring that the estimate of life expectancy at the age of the open interval is reasonable. Often the data on older people are scanty and particularly prone to errors. Thus, estimates of life expectancy based on these data can be implausible (usually too high).
- If completeness of reporting of deaths appears to be less than 60 per cent, then caution is advised in applying this method as the uncertainty about the estimate is large.

## APPLICATION OF METHOD

The method is applied in the following steps.

### *Step 1: Set the initial growth rate*

The growth rate can be estimated initially either from estimates of the total population above a certain age (chosen to best match the assumption that population is stable) at two time points or as estimated from the application of the Brass Growth Balance method. In the first instance, if one has estimates of the total population at time points $t_1$ and $t_2$, one would estimate the growth rate as follows:

$$r = \frac{\ln\left( {}_{\infty}N_x(t_2) / {}_{\infty}N_x(t_1)\right)}{t_2 - t_1}$$

where ${}_{\infty}N_x(t)$ is the population aged $x$ and older at time $t$.

### *Step 2: Estimate the life expectancy at age A and five-year age intervals down to 65*

This can be done in one of several ways.

1) Use estimates from an independent source if reliable estimates are available. Possible sources would be estimates produced by previous research or from population projections such as the World Population Prospects (UN Population Division 2011).
2) Use the estimates derived from the data after applying the Brass Growth Balance method. The workbook implementing that method produces such estimates as part of the output.
3) Use the ratio of the reported deaths in the age group 10 to 39 last birthday to those in the age group 40 to 59 last birthday (${}_{30}D_{10}/{}_{20}D_{40}$) to determine (by comparison) a level of the West model life table, from which estimates of life expectancy can be read. These estimates are included as part of the workbook implementing this method. Unfortunately, since the West model life table does not reflect mortality resulting from HIV/AIDS, this approach is unsuitable for countries that have significant numbers of AIDS deaths.
4) Solve for the life expectancy iteratively by starting with a reasonable guess such as one estimated from the West table (although in some cases this may not work in countries with significant numbers of AIDS deaths) or from an independent source. Then estimate completeness (as described below), copy the life expectancies from the *Life expectancies* spreadsheet of the associated workbook, paste the values into the *Method* spreadsheet of the associated workbook and re-estimate completeness. Repeat if necessary until the change to life expectancies is no longer significant. Unfortunately, if there are reasons for suspecting that, even after correcting the rates for incompleteness, mortality is underestimated at the older ages (for example, if there is significant age exaggeration, or relatively higher incompleteness at the older ages) this approach will overestimate the life expectancies and hence overestimate the overall level of completeness of reporting.

***Step 3: Estimate the number of people who turned x, and the number aged x to x + 4 last birthday, from the reported deaths***

The number of people who turned $x$ during the period over which the deaths were reported is estimated from the reported deaths as follows:

$$\hat{N}_x = \hat{N}_{x+5} \exp(5r) + {}_5D_x \exp(2.5r)$$

and

$$\hat{N}_A = {}_\infty D_A \left( \exp(r \times e_A) - (r \times e_A)^2 / 6 \right)$$

where $A$ is the age at the start of the open interval, $r$ is the annual population growth rate, and $e_A$ is the life expectancy at age $A$.

The number of people who were aged between $x$ and $x + 4$ last birthday during the period over which the deaths were reported is estimated from the numbers who turned $x$ in five-year intervals as follows:

$${}_5\hat{N}_x = 2.5\left( \hat{N}_x + \hat{N}_{x+5} \right).$$

***Step 4: Estimate the number of people who were aged x to x + 4 last birthday during the period over which the deaths are reported, from the census population***

The number of people who were aged $x$ to $x + 4$ last birthday during the period over which the deaths are reported is estimated from the census population by simply multiplying the numbers in the population in that age group by the length of the period over which the deaths are reported (measured in years).

***Step 5: Calculate the ratios of the estimates of the population aged x to x + 4 last birthday and the ratios of the population aged x to A–1 last birthday derived from deaths to those derived from the census population***

Two sets of ratios of the estimates derived from the deaths to those derived from the census population are calculated. The first is the ratios in quinquennial age groups, which are calculated directly. The second is the ratios of the numbers from age $x$ to the age of the open interval, $A$, with the numbers of people who turned $x$ to $A$–1 during the period being calculated as the aggregate of the numbers in five-year age groups between ages $x$ and $A$–5. In other words,

$${}_{A-x}\hat{N}_x = \sum_{a=x,5}^{A-5} {}_5\hat{N}_x.$$

***Step 6: Estimate the completeness of reporting of deaths***

In order to determine the level of completeness of reporting one first needs to decide if the initial choice of growth rate is correct. The interpretation of the plots of the ratios is discussed in more detail below. However, essentially the correct growth rate is identified as that which produces the most level set of ratios by age. The *Method* spreadsheet is set up so that *Solver* (Data, Solver, Solve) will find the growth rate that minimizes the absolute deviation from the mean of the ratios over the age range specified by the user.

If the initial estimate of the growth rate produces a level series of ratios across adult ages but with significant curvature downward at the older ages this could indicate a fall off in completeness at the older ages (as might be the case if, for example, people retired from urban areas to rural areas, where completeness of registration was lower). In such a situation it is important **not** to set the growth rate to produce a level set of ratios, but rather to use the initially chosen growth rate.

If one is also solving for the both growth rate and life expectancies iteratively, these values will need to be pasted from the *Life expectancies* spreadsheet into the *Method* spreadsheet and a new growth rate set. This process may need to be repeated two or three times, until there is no change in the life expectancies.

Finally, one decides on the age range of ratios to be used to determine the completeness. If there is a significant curvature upward at the older ages this probably indicates age exaggeration, particularly for deaths, and one needs to try and identify an age for the open interval below which the age exaggeration is not significant. If completeness drops off at ages below 35, this could indicate unaccounted for out-migration. If this is suspected then one should exclude these ages from determining the growth rate or completeness.

Completeness is estimated from the age group-specific ratios. In order to produce a robust estimate, it is calculated as the sum of 50 per cent of the median plus 25 per cent of each of the 75th and 25th percentile of these ratios.

However, since this is an estimate of the completeness on the assumption that the census population was at the mid-point of the period over which the deaths have been recorded, it is desirable to correct for any difference between the time of the census and the mid-point of the period over which the deaths were recorded. In order to do this we multiply this estimate of completeness by the ratio of the census population to the estimate of the population at

time $t_m$, on the assumption that the population, which is assumed to be stable, is growing at an annual growth rate estimated by $a$, i.e. $\exp\left(a\left(t_c - t_m\right)\right)$ where $t_c$ is the time of the census and $t_m$ is the mid-point of the period over which the deaths were recorded.

***Step 7: Estimate mortality rates adjusted for incompleteness of reporting of deaths***

In order to compute mortality rates one needs first to estimate the population in five-year age groups at the mid-point of the period over which the deaths were recorded by multiplying the census numbers by $\exp\left(-a\left(t_c - t_m\right)\right)$.

Next, one needs to adjust the number of deaths for incompleteness by dividing the reported number of deaths by the estimate of completeness, $c$.

The person-years of exposure are estimated by multiplying the estimated population as at $t_m$ by the length of the period over which the deaths were reported, $t$.

Mortality rates adjusted for the incompleteness of the reporting of deaths are thus estimated as follows:

$$ {}_5m_x = \frac{{}_5D_x / c}{t \times {}_5N_x \exp\left(-a\left(t_c - t_m\right)\right)}. $$

Since both the numerator (through the estimate of $c$), and the denominator are adjusted by $\exp\left(-a\left(t_c - t_m\right)\right)$, skipping these adjustments (in Steps 6 and 7) would still produce the same estimates of mortality rates. The estimate of completeness, however, would be equivalent to what it would be if the population at $t_m$ was assumed to be that at $t_c$.

***Step 8: Smooth using relational logit model life table***

Because the age-specific rates can be quite erratic they need to be graduated (smoothed). This can be achieved by fitting a Brass relational logit function to a sex-specific standard life table which is considered to have the same shape as that generated by the mortality in the population being investigated.

The workbooks (see website) contain a spreadsheet that allows one to produce a smooth set of mortality rates by using a relational logit model fitted to the life table generated by the adjusted mortality rates. The user can choose between the standard from the General family of United Nations model life tables or one from any of the four families of Princeton model life tables. A custom life table can be entered as standard if there is reason to assume that it better resembles the pattern of adult mortality in the population being studied.

In order to fit the model, probabilities of people aged $x$ dying in the next 5 years, ${}_5q_x$, are estimated from the adjusted rates of mortality as follows:

$$ {}_5q_x = \frac{5\,{}_5m_x}{1 + 2.5\,{}_5m_x}. $$

From this the life table with a radix of $l_5 = 1$ is calculated as follows:

$$ l_{x+5} = l_x\left(1 - {}_5q_x\right). $$

The coefficients, $\alpha$ and $\beta$ are determined by fitting the relational logit model as follows:

$$ \gamma_x = \alpha + \beta\gamma_x^s $$

where

$$ \gamma_x = 0.5 \ln\left(\frac{1 - l_x}{l_x}\right) $$

and the superscript $s$ designates values based on a standard life table.

The fitted life table is then generated from the standard life table using the coefficients $\alpha$ and $\beta$ as follows:

$$ \gamma_x^{fitted} = \alpha + \beta\gamma_x^s $$

and

$$ l_x^{fitted} = \frac{1}{\exp\left(1 + 2\gamma_x^{fitted}\right)}. $$

The smoothed mortality rates are derived from this life table as follows:

$$ {}_5m_x^{fitted} = \frac{l_x^{fitted} - l_{x+5}^{fitted}}{T_x - T_{x+5}} $$

and

$$ {}_\infty m_x^{fitted} = \frac{l_x^{fitted}}{T_x} $$

where

$$ T_x = \sum_{x=x,5}^{\omega} \frac{5}{2}\left(l_x^{fitted} + l_{x+5}^{fitted}\right) $$

i.e.

$$ T_x = T_{x+5} + \frac{5}{2}\left(l_x^{fitted} + l_{x+5}^{fitted}\right) $$

and $\omega$ is the age above which the life table has no more survivors.

The life expectancies are derived as follows:

$$ e_x = \frac{T_x}{l_x}. $$

In the case where one wants to estimate the life expectancies at the older ages iteratively, these values are then used to re-estimate the completeness.

## WORKED EXAMPLE

This example uses data on the numbers of women in the population from the El Salvadorian census in 1961 and on deaths from vital registration for the calendar year 1961. The example appears in the *PnC_El Salvador* workbook (see website). The reference date for the 1961 Census was midnight between 5 and 6 May, so the date of the census is entered as 06/05/1961 on the *Introduction* sheet.

### *Step 1: Set the initial growth rate*

The growth rate estimated using the population aged 10 and older from the 1950, 1961 and 1971 Censuses in *Manual X* is 2.8 per cent while that from the application of the Brass Growth Balance method to these data was 3.1 per cent, which is very close to the estimate derived from, as an example, the mid-year population estimates for 1955 and 1965 from the International Data Base of the US Census Bureau, as follows:

$$r = \frac{\ln\left(3017852/2221139\right)}{1965-1955} = 0.0307\,.$$

### *Step 2: Estimate the life expectancy at age A and five-year age intervals down to 65*

The estimates derived from the data after applying the Brass Growth Balance method are as shown in column 2 of Table 21.1.

The ratio of the reported deaths in the age group 10 to 39 last birthday (1706) to those in the age group 40 to 59 last birthday (1467) is $\frac{1706}{1467} = 1.16$. The life expectancies of the female West model life table which corresponds to this are determined (from the table in the *Life expectancies* spreadsheet of the workbook) by interpolation and are shown in column 3 of Table 21.1. For example for age 65:

$$e_{65} = 9.50 + \frac{1.171-1.16}{1.171-1.115}\left(9.86-9.50\right) = 9.552\,.$$

Solving for the life expectancy iteratively by starting with the estimates from the West table produces an estimate of the growth rate (as explained in more detail below) of 3.065 per cent and the final estimates of life expectancy which appear in column 4 of Table 21.1.

*TABLE 21.1 Life expectancies from different sources, females, El Salvador, 1961 Census*

| *x* | Brass Growth Balance | Princeton West | Iterative estimates |
|---|---|---|---|
| 65 | 13.4 | 9.55 | 13.1 |
| 70 | 10.4 | 7.38 | 10.2 |
| 75 | 7.9 | 5.57 | 7.8 |
| 80 | 5.9 | 4.06 | 5.8 |
| 85 | 4.4 | 2.88 | 4.3 |

Since HIV/AIDS was not an issue in El Salvador back in 1961, one could use the estimates derived from the West life tables given in the *Life expectancies* spreadsheet of the workbook (see website) to estimate the completeness of reporting of deaths. However, for illustrative purposes the workbook has used the iterative estimates, even though comparison of the estimates in Table 21.1 (and of the observed mortality rate for the open age interval 75+ with that of the graduated rates) suggests that there is either age exaggeration or a fall-off in completeness in the data above age 75 which is likely to lead to a slight overestimate in completeness.

### *Step 3: Estimate the number of people who turned x, and those aged x to x+4 last birthday, from the reported deaths*

The number of people who turned *x* during the period over which the deaths were reported as estimated from the numbers of deaths in each age group using an open interval of 75+, the growth rate of 3.065 per cent and the estimate of life expectancy given in the fourth column of Table 21.1, are as shown in column 4 of Table 21.2. For example, the estimate of the number of people who turned 70 in the period over which the deaths were reported is calculated as follows:

$$\hat{N}_{75} = 1360\begin{pmatrix}\exp\left(.03065\times 7.76\right)\\ -\left(0.03065\times 7.76\right)^2/6\end{pmatrix} = 1712.11$$

$$\hat{N}_{70} = 1712\exp\left(5\times 0.03065\right) + 504\exp\left(2.5\times 0.03065\right) = 2539.76\,.$$

The number of people aged *x* to *x*+4 last birthday during the period over which the deaths were reported, estimated from the reported deaths is given in column 5 of Table 21.2. For example, the number who turned 20 to 24 last birthday is calculated as follows:

$${}_5\hat{N}_{20} = 2.5\left(21542+18212\right) = 99382.90\,.$$

***Step 4: Estimate the number of people who were aged x to x+4 last birthday during the period over which the deaths are reported, from the census population***

As the deaths are recorded over a single year the number of people who aged $x$ to $x+4$ last birthday during the period over which the deaths were reported is the number in the census for that age group (i.e. the numbers in column 6 are the same as those in column 2 of Table 21.2) as multiplication by one leaves the numbers unchanged.

***Step 5: Calculate the ratios of the estimates derived from deaths to those derived from the census population***

The ratios of the numbers of people aged $x$ to $x+4$ last birthday during the period over which the deaths were reported estimated from the reported deaths to those estimated from the census are given in columns 7 and 8 of Table 21.2. Examples of these calculations for age 65 are as follows:

$$\frac{{}_5\hat{N}_{65}}{{}_5N_{65}} = \frac{14962}{14964} = 0.9999$$

$$\frac{{}_{10}\hat{N}_{65}}{{}_{10}N_{65}} = \frac{14962+10630}{14964+11205} = 0.9779.$$

***Step 6: Estimate the completeness of reporting of deaths***

Although the estimate of the growth rate produced by the application of the Brass Growth Balance method produces a satisfactorily level series of ratios, for this example it was decided, for illustrative purposes, to solve for the growth rate and determine the life expectancies iteratively. This produced the plot of ratios shown in Figure 21.1.

Since there is no consistent trend (either upwards or downwards) apparent in Figure 21.1 the growth rate was determined using data for ages 5 to 74 by minimizing the deviations from the mean using *Solver*. Completeness was estimated from the ratios in the age range 15 to 64 to avoid the fluctuations in the estimates for completeness at the oldest ages (although this aspect of determining the estimate is fairly robust to fluctuations at individual ages). This produced an estimate of completeness of 89 per cent as follows:

$$\begin{aligned} c &= \left(0.5\times 0.8764 + 0.25\left(0.8575 + 0.9144\right)\right) \\ &\quad \exp\left(0.0306\left(1961.34 - 1961.50\right)\right) \\ &= 0.8812\times 1.005 = 88.6\%. \end{aligned}$$

*TABLE 21.2 Calculation of the numbers of people aged x to x+4 from the reported deaths and from the census and the ratios of the estimates, El Salvador, 1961 Census*

| Age | ${}_5N_x(t_c)$ | ${}_5D_x$ | Est $N_x$ | Est ${}_5N_x$ | Obs ${}_5N_x$ | $c$: ${}_5N_x$ | $c$: ${}_{A-x}N_x$ |
|---|---|---|---|---|---|---|---|
| 0–4 | 214,089 | 6,909 | | | 214,089 | | |
| 5–9 | 190,234 | 610 | 35,431 | 163,158 | 190,234 | 0.8577 | 0.8879 |
| 10–14 | 149,538 | 214 | 29,832 | 138,071 | 149,538 | 0.9233 | 0.8946 |
| 15–19 | 125,040 | 266 | 25,396 | 117,344 | 125,040 | 0.9384 | 0.8885 |
| 20–24 | 113,490 | 291 | 21,542 | 99,383 | 113,490 | 0.8757 | 0.8778 |
| 25–29 | 91,663 | 271 | 18,212 | 83,962 | 91,663 | 0.9160 | 0.8783 |
| 30–34 | 77,711 | 315 | 15,373 | 70,677 | 77,711 | 0.9095 | 0.8690 |
| 35–39 | 72,936 | 349 | 12,897 | 59,098 | 72,936 | 0.8103 | 0.8584 |
| 40–44 | 56,942 | 338 | 10,742 | 49,112 | 56,942 | 0.8625 | 0.8741 |
| 45–49 | 46,205 | 357 | 8,903 | 40,525 | 46,205 | 0.8771 | 0.8781 |
| 50–54 | 38,616 | 385 | 7,307 | 33,049 | 38,616 | 0.8558 | 0.8785 |
| 55–59 | 26,154 | 387 | 5,913 | 26,567 | 26,154 | 1.0158 | 0.8892 |
| 60–64 | 29,273 | 647 | 4,714 | 20,398 | 29,273 | 0.6968 | 0.8295 |
| 65–69 | 14,964 | 449 | 3,445 | 14,962 | 14,964 | 0.9999 | 0.9779 |
| 70–74 | 11,205 | 504 | 2,540 | 10,630 | 11,205 | 0.9487 | 0.9487 |
| 75+ | 16,193 | 1,360 | | | | | |

***Step 7: Estimate mortality rates adjusted for incompleteness of reporting of deaths***

The population as at the mid-point of the period over which the deaths were recorded is estimated by adjusting the census population for the growth between the two dates at the estimated growth rate of 3.1 per cent. These estimates are shown in the second column of Table 21.3. For example, for the 15–19 age group the number is estimated as follows:

$$
\begin{aligned}
{}_5N_{15}(t_m) &= 125040 \times \exp\left(-0.0306\left(1961.34 - 1961.50\right)\right) \\
&= 125040e^{0.0306\times 0.16} = 125655.7.
\end{aligned}
$$

Next the deaths are adjusted for incompleteness by dividing the number of reported deaths in each age group by the estimate of completeness. These numbers are shown in column 3 of Table 21.3. For example, for the 15–19 age group the number is derived from the number of reported deaths (shown in column 3 of Table 21.1), 266, as follows:

$$\frac{266}{0.8855} = 300.4.$$

The adjusted person-years of life lived (column 4 of Table 21.3) are the numbers in the population at the mid-point of the period over which the deaths have been recorded (column 2 Table 21.3) multiplied by the length (in years) of the period over which the deaths are recorded, which in this case is 1 year.

The mortality rates adjusted for incompleteness of reporting of deaths (column 5 of Table 21.3) are derived by dividing the adjusted deaths by the adjusted person-years of life lived. For example, for the 15–19 age group the adjusted rate is calculated as follows:

$$\frac{300.40}{125656} = 0.00239.$$

***Step 8: Smooth using relational logit model life table***

Estimates of probabilities of women aged $x$ dying in the next 5 years, ${}_5q_x$, estimated from the adjusted rates of mortality are shown in the second column of Table 21.4. For example, the probability of a 15-year old woman dying before reaching age 20 is calculated as follows:

$${}_5q_{15} = \frac{5\times 0.00239}{1+2.5\times 0.00239} = 0.0119.$$

The life table proportions of five-year olds alive at age $x+5$ estimated from the proportion alive at age $x$ using these values appear in column 3 of Table 21.4. For example the proportion alive at age 20 is calculated as follows:

***FIGURE 21.1*** *Completeness by age of registered deaths, El Salvador, 1961 Census*

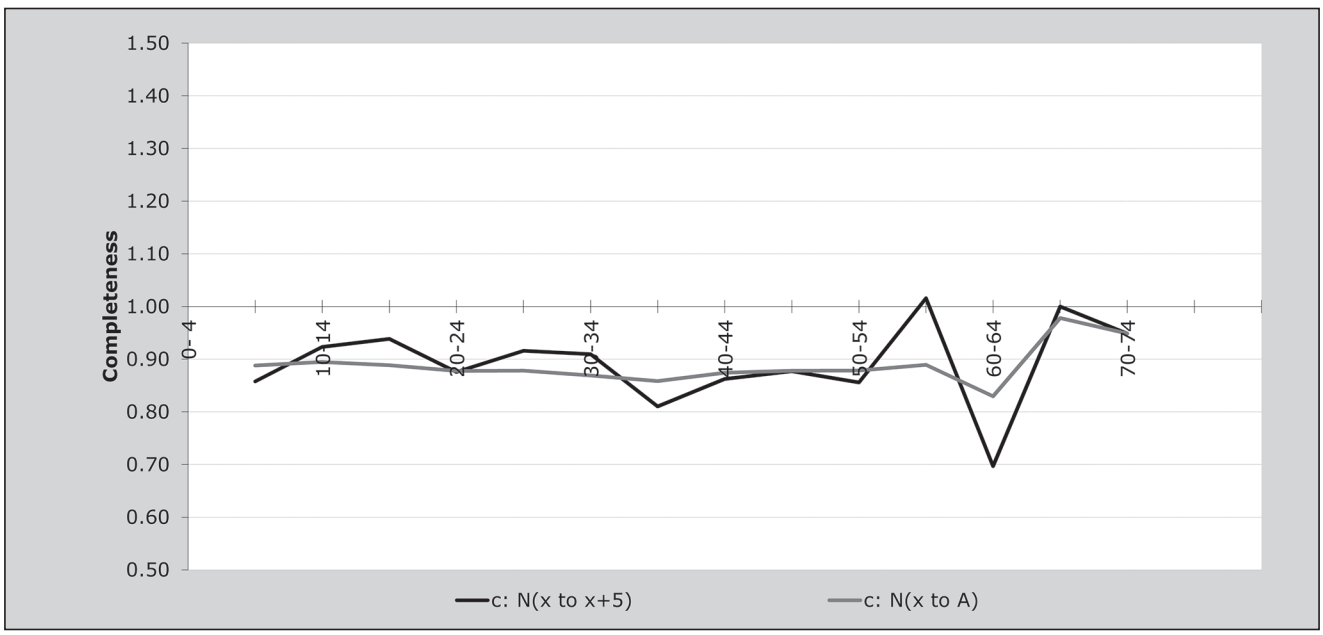


*TABLE 21.3 Calculation of adjusted mortality rates, El Salvador, 1961 Census*

| Age | Adjusted ${}_5N_x(t_m)$ | Adjusted ${}_5D_x$ | Adjusted $PYL(x,5)$ | Adjusted ${}_5m_x$ |
|---|---|---|---|---|
| 0–4 | | | | |
| 5–9 | 191,171 | 689 | 191,171 | 0.0036 |
| 10–14 | 150,274 | 242 | 150,274 | 0.0016 |
| 15–19 | 125,656 | 300 | 125,656 | 0.0024 |
| 20–24 | 114,049 | 329 | 114,049 | 0.0029 |
| 25–29 | 92,114 | 306 | 92,114 | 0.0033 |
| 30–34 | 78,094 | 356 | 78,094 | 0.0046 |
| 35–39 | 73,295 | 394 | 73,295 | 0.0054 |
| 40–44 | 57,222 | 382 | 57,222 | 0.0067 |
| 45–49 | 46,433 | 403 | 46,433 | 0.0087 |
| 50–54 | 38,806 | 435 | 38,806 | 0.0112 |
| 55–59 | 26,283 | 437 | 26,283 | 0.0166 |
| 60–64 | 29,417 | 731 | 29,417 | 0.0248 |
| 65–69 | 15,038 | 507 | 15,038 | 0.0337 |
| 70–74 | 11,260 | 569 | 11,260 | 0.0505 |
| 75+ | 16,273 | 1,536 | 16,273 | 0.0944 |

$$l_{20} = 0.9743\left(1 - 0.01188\right) = 0.9627.$$

The logit transformations of the proportions surviving appear in column 4 of Table 21.4. For example, the logit transformation of the $l_{20}$ is calculated as follows:

$$\gamma_{20} = 0.5\ln\left(\frac{1-0.9627}{0.9627}\right) = -1.6254\,.$$

The logit transformation of the conditional life table for females based on the West family of Princeton model life tables with $e_0 = 60$ in column 5 of Table 21.4 appears in column 6 of Table 21.4. As can be seen from Figure 21.2 the West model appears to fit the data well, with the possible exception of the youngest ages.

The coefficients, $\alpha$ and $\beta$ are determined as the intercept and slope of the straight line fitted to the logit transformations in columns 4 and 6 of Table 21.4 over the range of ages chosen by the user (between 45 and 75 in this example), namely 0.0094 and 0.9754 respectively. The range 45 to 75 is chosen because the fit to the older ages is of importance

*TABLE 21.4 Calculation of smoothed mortality rates using a relational logit model life table, El Salvador, 1961 Census*

| Age | ${}_5q_x$ | $l_x/l_5$ | Obs. $Y(x)$ | Princeton West Cdn. $l^s(x)$ | Cdn. $Y^s(x)$ | Fitted $Y(x)$ | Fitted $l(x)$ | $T(x)$ | $e(x)$ | Smooth ${}_5m_x$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | | | | | | | | | |
| 5 | 0.0179 | 1 | | 1.0000 | | | 1 | 61.383 | 61.4 | 0.0025 |
| 10 | 0.0080 | 0.9821 | –2.0037 | 0.9890 | –2.2506 | –2.1860 | 0.9875 | 56.414 | 57.1 | 0.0019 |
| 15 | 0.0119 | 0.9743 | –1.8172 | 0.9805 | –1.9585 | –1.9011 | 0.9782 | 51.500 | 52.6 | 0.0028 |
| 20 | 0.0143 | 0.9627 | –1.6254 | 0.9681 | –1.7060 | –1.6547 | 0.9648 | 46.643 | 48.3 | 0.0036 |
| 25 | 0.0165 | 0.9489 | –1.4611 | 0.9519 | –1.4928 | –1.4468 | 0.9475 | 41.862 | 44.2 | 0.0041 |
| 30 | 0.0225 | 0.9333 | –1.3192 | 0.9337 | –1.3226 | –1.2808 | 0.9284 | 37.173 | 40.0 | 0.0047 |
| 35 | 0.0265 | 0.9123 | –1.1709 | 0.9132 | –1.1766 | –1.1384 | 0.9069 | 32.584 | 35.9 | 0.0054 |
| 40 | 0.0328 | 0.8881 | –1.0356 | 0.8899 | –1.0447 | –1.0097 | 0.8828 | 28.110 | 31.8 | 0.0064 |
| 45 | 0.0425 | 0.8589 | –0.9033 | 0.8628 | –0.9194 | –0.8875 | 0.8551 | 23.765 | 27.8 | 0.0080 |
| 50 | 0.0545 | 0.8224 | –0.7665 | 0.8299 | –0.7925 | –0.7637 | 0.8216 | 19.573 | 23.8 | 0.0110 |
| 55 | 0.0798 | 0.7776 | –0.6259 | 0.7863 | –0.6514 | –0.6261 | 0.7777 | 15.575 | 20.0 | 0.0153 |
| 60 | 0.1169 | 0.7156 | –0.4613 | 0.7289 | –0.4946 | –0.4731 | 0.7204 | 11.830 | 16.4 | 0.0232 |
| 65 | 0.1555 | 0.6319 | –0.2702 | 0.6490 | –0.3074 | –0.2905 | 0.6413 | 8.426 | 13.1 | 0.0354 |
| 70 | 0.2244 | 0.5336 | –0.0674 | 0.5427 | –0.0856 | –0.0742 | 0.5370 | 5.480 | 10.2 | 0.0566 |
| 75 | #N/A | 0.4139 | 0.1739 | 0.4062 | 0.1898 | 0.1945 | 0.4039 | 3.128 | 7.7 | 0.0897 |
| 80 | #N/A | #N/A | #N/A | 0.2545 | 0.5373 | 0.5335 | 0.2560 | 1.478 | 5.8 | 0.1399 |
| 85 | #N/A | #N/A | #N/A | 0.1201 | 0.9956 | 0.9805 | 0.1234 | 0.530 | 4.3 | 0.2109 |

for estimating the life expectancy at the age of the start of the open interval.

These coefficients are then applied to the logit transformation of the conditional model life table to produce the fitted logits in column 7 of Table 21.4. Thus, for example the fitted logit at age 20 is calculated as follows:

$$\gamma_{20}^{fitted} = 0.0094 + 0.9754 \times (-1.7060) = -1.6547.$$

These values are then used to produce the fitted life table in column 8 of Table 21.4. For example the value at age 20 is calculated as follows:

$$l_{20}^{fitted} = \frac{1}{\exp(1 + 2 \times (-1.6547))} = 0.9648.$$

The conditional years of life lived, $T_x$, which appear in column 9 of Table 21.4 are then calculated from the fitted life table. These numbers are used to produce the smoothed mortality rates which appear in column 11 of Table 21.4. For example, for age 80

$$T_{80} = 0.53 + \frac{5}{2}(0.256 + 0.1234) = 1.478$$

$$_5m_{80}^{fitted} = \frac{.256 - 0.1234}{1.478 - 0.53} = 0.1399.$$

The life expectancies which appear in column 10 of Table 21.4 are the numbers in column 9 divided by the numbers in column 8. For example, the life expectancy at age 65 is

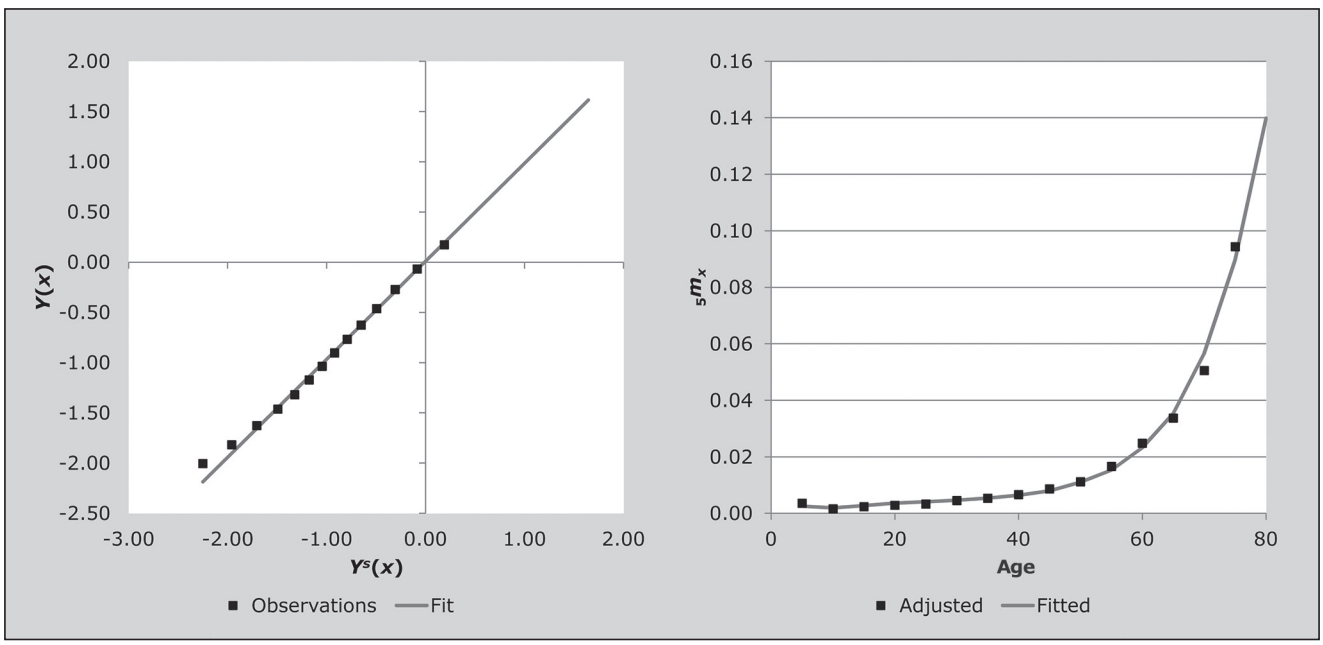
$$\frac{8.426}{0.6413} = 13.139.$$

## DIAGNOSTICS, ANALYSIS AND INTERPRETATION

### Checks and validation

The example above was taken from *Manual X* (UN Population Division 1983) which produced an estimate of completeness of around 83 per cent from applications of both this method and the Brass Growth Balance method. The difference between the two estimates in *Manual X* and the one produced in this application (89 per cent) appears to be largely due to differences in the method of estimating the population at the age of the open interval (*A*). The full effect is counteracted to some extent by a reduction (relative to the estimate in *Manual X*) due to the fact that the current approach solved for a growth rate that was higher (3.02 per

*FIGURE 21.2 Observed logits and adjusted mortality rates against expected derived from the female West model life table, El Salvador, 1961 Census*



cent) than that used in *Manual X* (2.87 per cent). Thus applying the method using the estimates of life expectancy calculated from the West model life table (column 3, Table 21.1) produced an estimate of completeness of 85 per cent.

**Interpretation**

As there is no consistent trend (upward or downward) in the plotted series in Figure 21.1, there is no reason to reduce the age of the open interval. However, had this been necessary, it would have created a problem in deciding which estimate of completeness to accept, since the estimate of completeness for an open-ended age interval of 70+ is 85 per cent, while that for an interval of 65+ is 76 per cent. The spreadsheet does not allow the open interval to be less than 65+, but, had one used an interval of 60+, the estimate of completeness would have been higher than 76 per cent. As a general rule, it is not recommended in a population with significant digital preference to truncate at an age ending in zero.

Taken together, these estimates suggest that the completeness of death reporting is about 85 per cent, somewhat lower than the 92 per cent estimated by applying the Brass Growth Balance method to these data. Interestingly, had one used the estimates of life expectancy derived from the West life table on the basis of the ratio of ${}_{30}D_{10}$ to ${}_{20}D_{40}$ (and an open interval of 75+) the estimate of completeness drops to 85 per cent but the life expectancy derived from the smoothed rates is closer to that derived iteratively than to those used to produce the estimate of completeness. This suggests that the method is not very sensitive to the estimate of life expectancy used, particularly if the open interval starts at a high age.

## METHOD-SPECIFIC ISSUES WITH INTERPRETATION

**Source of reported deaths**

Generally there are two sorts of problems with death data: those that lead to under/over coverage that is constant by age, which is precisely what the method is intended to address, and those which lead to differential coverage by age, which can distort the estimates. Although the general approach remains essentially the same irrespective of the source of the death data, different sources of data are prone to different biases which might impact on the interpretation of the results. These are illustrated by way of particular examples, but, in general terms, you need to look out for the following biases in the death data.

***1) Vital registration***

If the proportionate split of the population between urban and rural (or appropriate proxies) areas differs significantly by age and the completeness of reporting of deaths in urban areas is significantly higher than it is in rural areas, then the assumption that completeness is independent of age is likely to be violated by a falling off of completeness with age at ages over 50 if a proportion of people move from urban to rural areas on retirement. If ignored and the growth rate is estimated using *Solver*, this violation is likely to lead to an underestimate of the average level of completeness.

***2) Deaths reported by households in censuses/surveys***

The data are subject to three potential problems:

- If a significant proportion of households dissolve on the death of a key person (e.g. the sole breadwinner), then the deaths of such people go unreported, leading to a violation of the assumption that completeness is invariant with age. If a significant proportion of deaths in some age groups are of individuals who do not live in private households (for example, they live in homes for the elderly), the breach of the assumption could be even more severe. However, this is not an issue in most developing countries.
- In situations where young adults leave the home they grew up in to work in urban areas, it is possible that they are regarded as being members of more than one household (or of neither household) and their deaths could be reported more than once (or not at all), again leading to a violation of the assumption of constant reporting of deaths by age. In this case, one can limit the impact by ignoring the data below a specific age in determining completeness.
- Reference period error: Since there is often confusion about the exact period for which deaths are to be reported, in addition to uncertainty about exact dates of death, it is possible for there to be overall under- or over-reporting of deaths. Provided one can assume that this is independent of the age of the deceased, this distortion will be accounted for in the estimate of completeness and is not a problem for estimating mortality rates.

***3) Deaths recorded in health facilities***

Little is known about how well this source of data works. However, it can be expected that completeness would depend on the distribution of health services from which the data have been gathered, and in many developing countries such services are likely to be concentrated in urban areas. So

again, if the proportion of the population living in urban rather than rural areas varies with age, then completeness cannot be assumed to be independent of age. It is also possible that certain causes will predominate in facilities, and if these causes are significant and age-related, this could lead to a further violation of the assumption of constant completeness by age.

In all such cases, one should avoid the temptation of adjusting the growth rate to produce a level sequence of the ratios. Instead one should ensure that the estimate of $c$ is determined over a range of ages that excludes those in which death reporting is either exceptionally complete or exceptionally incomplete.

**General diagnostic interpretation**

In practice the sequences of both ${}_5\hat{N}_x \big/ {}_5N_x$ and ${}_A\hat{N}_x \big/ {}_AN_x$ are affected by violations of the assumptions. However, part of the power of this technique is that most of the typical violations of assumptions produce fairly distinctive characteristic deviations from the expected horizontal line and in certain circumstances these patterns are interpretable. The following are examples:

- *Incorrect growth rate*: If $r$ is too high the sequences of points fall nearly linearly with increasing age towards the underlying value of completeness and vice versa, as can be concluded from inspection of Equation (1) below. The effect is greater for ${}_5\hat{N}_x \big/ {}_5N_x$ than for ${}_A\hat{N}_x \big/ {}_AN_x$.
- *Exaggeration of reported age*: Typically, relatives reporting deaths exaggerate the person's age at death more than living individuals reporting their own ages. This produces rising sequences of points which are imperceptible up to the age at which exaggeration begins, followed by a sharp upward curve thereafter. Again it can be seen from inspection of Equation (1) below in that age exaggeration not only leads to an increase in the number of deaths in the older age categories, but, in addition, transfers within a category lead to those deaths being multiplied by a larger exponential term, although this effect is smaller. Such a pattern would also be produced by rising completeness in death registration with age above a certain age. However, there appears to be no evidence of this in practice (Preston, Coale, Trussell *et al.* 1980).
- *Age misstatement in the population estimates and age-specific miscounting*: This is exhibited by an erratic sequence of the ratios over the age span. Since ${}_5\hat{N}_x$ is cumulative in form, it tends to follow the age distribution of the population quite closely. Thus if there are zigzags it is likely that the peaks may be associated with age aversion or under-enumeration in the population and troughs with age heaping or over-enumeration in the population. If these fluctuations are independent of the age, they should not distort the estimate of completeness particularly. Blacker (1988) suggested using age groups 18–22, 23–27, etc. to remove zigzags and showed that for the Brass Growth Balance method this removed bias in the estimate of the slope. However, if these distortions are systematic, e.g. unaccounted for migration below a certain age, it may be better to exclude these points from estimating the completeness.

Generally the effect of overstated ages can be largely removed by beginning the open interval at a sufficiently young age to confine most of the overstatement to the open interval.

In order to distinguish a declining sequence of ratios due to improving mortality from that due to the choice of too high a growth rate, one needs to look to evidence from other sources to determine which the more likely explanation is. If the population has experienced a decline in mortality, the median of the ratios of cumulated populations from 10 to, say, 45 ought still to provide a reasonable estimate of the completeness of death registration. Although this method has a lot to recommend it, and is more robust to departure from stability than the Brass Growth Balance method, it is more sensitive than the latter to certain types of age misreporting. Thus, it will not always be possible to obtain a single robust estimate of the completeness of the death data unless one can confirm the assumptions (particularly the growth assumption) by other means.

## DETAILED DESCRIPTION OF METHOD

**Mathematical exposition**

The Preston and Coale method is a special case of the Synthetic Extinct Generations method, with the growth rate of the population aged $x+$, $r(x+)$ constant for all ages.

The method arises out of work by Preston and Hill (1980) further developed by Preston, Coale, Trussell *et al.* (1980) and has its origins in the method of extinct generations originally proposed by Vincent (1951). It is based on the idea that the number of persons at a particular age at a point in time must equal the total number of deaths arising from this cohort from that time until the last survivor has died.

In a stable and closed population the relationship is:

$$N_x = \sum_{a=x}^{\varpi} D_a e^{r(a-x)} \quad (1)$$

where $D_a$ are the deaths at the same point in time as $N_x$ since in a stable closed population $D_{a,t}$, the deaths aged $a$ which are expected to occur $t$ years from the year for which we have recorded deaths, is equal to $D_a e^{rt}$.

If instead of $D_a$ we know $D_a^r$, the recorded number of deaths aged $x$ last birthday, and if we estimate the population aged $x$, $\hat{N}_x$, by $\hat{N}_x = \sum_{a=x}^{\varpi} D_a^r e^{r(a-x)}$ then $\hat{N}_x / N_x$, where $N_x$ is the true population at the mid-point of the period over which the deaths have been recorded, gives an indication of the percentage registration for ages $x$ and over, $c_{x+}$. If the $N_x$ are available at some other point in time, then they can be adjusted for the growth over the period between the two times using the growth rate $r$. However, if the level of completeness is being estimated in order to calculate mortality rates, the same correction would, in effect, be made to both the numerator and the denominator and thus could be ignored.

There is, however, a problem in computing $\hat{N}_x$ in practice in that the $D_a^r$ are unlikely to be available beyond a certain age (and even if they are, are unlikely to be very accurate) with all reported deaths above that age being grouped together in an open interval, $D_{A+}^r$ where $A$ is the lower bound of the age interval. However, various methods have been suggested to deal with this problem. For example, *Manual X* (UN Population Division 1983: 134) suggests that by assuming that the pattern of mortality fits one of the Princeton model life tables (Coale, Demeny and Vaughan 1983), $\hat{N}_A$ can be estimated as follows:

$$\hat{N}_A = D_{A+} e^{rz(A)}$$

where

$$z(A) = a(A) + b(A) + c(A) e^{\left[D_{45+}/D_{10+}\right]}.$$

The coefficients have been tabulated (Table 123, UN (1983: 134, 134)) and $D_{45+}/D_{10+}$ is estimated by $D_{45+}^r / D_{10+}^r$.

Alternatively, Bennett and Horiuchi (1984) suggested that the population aged $A$ can be estimated using the following formula:

$$\hat{N}_A = {}_{\infty}D_A \left( \exp\left(r \times e_A\right) - \left(r \times e_A\right)^2 / 6 \right)$$

where the life expectancy is interpolated from the West family of Princeton model life tables on the basis of the ratio of the reported deaths between ages of 10 and 40 to those between ages 40 and 60.

Since $\hat{N}_x$ can be approximated by

$$N_{x+5} e^{5r} + {}_5 D_x^r e^{2.5r}$$

once $\hat{N}_A$ has been estimated the $\hat{N}_x$ can be estimated from the ${}_5 D_x^r$.

**Limitations**

The major limitations of the method as described above and provided for in the spreadsheet are that it requires that the population be stable and closed to migration and it should not be applied when these conditions do not apply to any significant extent. By way of example of inappropriate usage, application of this method (data available in the should be *SEG_South Africa_males* workbook) to estimate completeness of reporting of deaths in South Africa between the 2001 Census and a census replacement survey in 2007, estimating the population in the middle of the period as that average of the two survey populations, provides an estimate of completeness, using the same age range, of 84 per cent. Increasing the minimum age of the range of the data used to fit the straight line to 35 increases the estimate to 86 per cent, still somewhat lower than the estimate of 94 per cent produced using the Synthetic Extinct Generations method.

This method is more vulnerable to age misreporting than the Brass Growth Balance method. In particular, as mentioned above, the common tendency to exaggerate the age reported at death (relative to that recorded at census) will manifest itself by the plotted points rising noticeably from the age above which the ages have been exaggerated. In such a situation it is better to use the growth rate estimated by the Brass Growth Balance method. In addition the method is also, as demonstrated above, sensitive to the choice of open interval if there is extreme digit preference in the data. This is most likely with census data.

The method is less vulnerable to the effects of destabilization resulting from a rapid change in mortality (Martin 1980). However, as simulation has shown for the Brass Growth Balance method (Rashad 1978), the bias resulting from a slow steady improvement in mortality (as has been experienced by some developing countries in the absence of epidemics, famine and wars) is quite small.

As far as changes in fertility rates are concerned these tend to have little impact on the performance of the method since they affect mainly the youngest age groups, which have a limited influence on the estimate of completeness. If necessary, these age groups can be excluded from determining the growth rate and estimate of completeness.

Migration is likely to affect the young adult population (mainly between 20 and 35) but to have much less effect on deaths, which occur largely in old age. Unaccounted-for immigration will tend to lower the slope and hence lead to an over-estimate of the extent of death registration and an underestimate of mortality rates. Unaccounted-for emigration will have the opposite effect. Some demographers advocate fitting the straight line to data down to age 5 to limit the effect of unaccounted-for migration, on the assumption that any differences in completeness of reporting of deaths at these younger ages from that of the older ages is unlikely to lead to any major distortions since the mortality is very light between ages 5 and 14. However, it doubtful that this adaptation removes much of the bias.

Alternatively one could confine the fit to points above age 35 to remove the bulk of the effect of migration. However, often the data at the older ages are more suspect making the estimate of completeness less reliable. Although using these adaptations probably produces better estimates than simply ignoring migration, there is, unfortunately, little research into the accuracy of the estimated completeness produced by these adaptations.

Technically, if one had reliable estimates of net migration by age, one could adapt the method by replacing the growth rate $r$ by $r-{}_5i_x$, where ${}_5i_x$ is the net in-migration rate for the age group $x$ to $x+4$ last birthday, in deriving $\hat{N}_x$. However, in practice, in situations where one has to apply this method one rarely has sufficiently reliable estimates of net migration by age to warrant adapting the method.

Fluctuations in the completeness of death registration with age are likely to introduce curvature in the pattern of points. Consequently, it is one of the strengths of this method that if the points for successive age boundaries fall on a reasonably level line then it is probably reasonable to assume that completeness is constant with respect to age. However, where some but not all the points lie on a straight line one may decide to limit the age range used to determine the estimate of completeness.

## FURTHER READING AND REFERENCES

Since this method is a particular case of the more general Synthetic Extinct Generations method, readers are referred to Chapter 25 for further reading.


Bennett NG and S Horiuchi. 1984. "Mortality estimation from registered deaths in less developed countries", *Demography* **21**(2):217–233. doi: 10.2307/2061041

Blacker J. 1988. *An Evaluation of the Pakistan Demographic Survey*. Karachi: Pakistan Federal Bureau of Statistics.

Coale AJ, P Demeny and B Vaughan. 1983. *Regional Model Life Tables and Stable Populations.* New York: Academic Press.

Martin LG. 1980. "A modification for use in destabilized populations of Brass's Technique for estimating completeness of death registration", *Population Studies* **34**:381–395. doi: 10.2307/2175194

Preston SH, AJ Coale, J Trussell and M Weinstein. 1980. "Estimating the completeness of reporting of adult deaths in populations that are approximately stable", *Population Index* **46**:179–202. doi: 10.2307/2736122

Preston SH and K Hill. 1980. "Estimating the completeness of death registration", *Population Studies* **34**:394–366. doi: 10.2307/2175192

Rashad HM. 1978. "The Estimation of Adult Mortality from Defective Registration Data." Unpublished PhD thesis, London: University of London.

UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

UN Population Division. 2011. *World Population Prospects: The 2010 Revision, Volume I: Comprehensive Tables*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/313. http://esa.un.org/unpd/wpp/Documentation/pdf/WPP2010_Volume-I_Comprehensive-Tables.pdf

Vincent P. 1951. "La mortalite des vieillards", *Population* **6**:182–204. doi: 10.2307/1524149

# Chapter 22 Indirect estimation of adult mortality from orphanhood

*Ian M Timæus*

## DESCRIPTION OF THE METHOD

Orphanhood methods estimate the mortality of adult women and men indirectly from data on the survival status of respondents' mothers and fathers respectively. In order to apply the method, at least one census or single-round survey of the population must have included the questions 'Is your mother alive?' and 'Is your father alive?'. Mortality can be estimated from the answers to these questions without requiring respondents to recall the dates when deaths occurred or the ages at death of deceased individuals.

Since respondents' mothers must have been alive when the respondents were born, the duration over which they have been exposed to the risk of dying equals the age of the respondents. By allowing for the mean age at which the mothers gave birth in the population concerned, it is possible to predict life table survivorship from age 25 to age 25 plus a rounded number of years ($n$) based on the age group of the respondents ($l_{25+n}/l_{25}$) from the proportion of respondents in each age group whose mother is alive. Similarly, by adjusting for the mean age at which the fathers have children, one can predict life table survivorship of adult men from the proportions of respondents with living fathers. As men tend to be older than their wives and other partners, their survivorship is measured between a base age of 35 and age $35 + n$ where $n$ is again linked to the age group of the respondents.

If mortality has changed over time, the estimated survivorship ratios reflect the mortality rates that have prevailed at a range of ages and dates. A 'time location' method has been developed that estimates how many years prior to inquiry each cohort survivorship ratio equalled the period survivorship ratio. These intervals increase with the age of respondents, ranging between about 4 and 14 years before the collection of the data. Thus, if the survivorship ratios estimated from the reports of different age groups of respondent are translated into a common index of mortality in adulthood (such as ${}_{45}q_{15}$) using a 1-parameter system of model life tables, these statistics will refer to different dates and can be used to infer the broad trend in mortality over time.

One advantage that orphanhood methods have over questions about household deaths is that only censuses or unusually large surveys can capture information on enough deaths in households in the year before the inquiry to yield mortality estimates that are sufficiently precise to be useful. The orphanhood method can be used in much smaller inquiries, although all methods for the estimation of adult mortality require data on thousands of households. Moreover, the method does not assume that the population is closed to migration. However, the results from the method will not be representative for small states or sub-national areas in which a substantial proportion of the population are in-migrants or have emigrated.

## DATA REQUIREMENTS AND ASSUMPTIONS

### Tabulations of data required

To estimate the mortality of adult women:

- The proportion of respondents whose mother is alive by five-year age group of respondent. (Those who did not know or did not declare their mother's survival status should be excluded from the calculations.)
- The number of births in the year before a demographic inquiry tabulated by five-year age group of the women giving birth.

To estimate the mortality of adult men:

- The proportion of respondents whose father is alive by five-year age group of respondent. (Those who did not know or did not declare their father's survival status should be excluded from the calculations.)

- The number of births in the year before a demographic inquiry tabulated by five-year age group of the women giving birth.
- An estimate of the difference between the ages of men and women having children, such as the difference between the median ages of currently married men and women.

These tables should generally be produced separately for male and female respondents and estimates made from both sets of proportions and for the two sexes combined.

For the purposes of estimating mortality, ideally everyone aged less than 50 should be asked the questions about the survival of their mothers and fathers (it may be simpler just to ask everyone these questions). Nevertheless, if an inquiry only asked about the orphanhood of children, the data can be used to produce point estimates of women's and men's mortality a few years earlier.

If sample or design weights have been provided with the data, remember to apply them in the manner appropriate to your statistical software when deriving the tabulations used as inputs.

**Important assumptions**

An inherent limitation of the orphanhood method is that data on parents' survival can only be collected from those of their offspring who are alive themselves. The survival of adults who have no living children is unrepresented in the reported proportions of parents alive. Moreover, parents with more than one surviving child are over-represented in comparison to those with exactly one surviving child in proportion to the number of their surviving children. Thus, the method only produces unbiased results if the mortality of the parents is unrelated to how many of their children are alive at the time that the data are collected. In general, the selection bias that arises from breaches in this assumption is small (Palloni, Massagli and Marcotte 1984). In populations affected by generalized HIV epidemics, however, it is likely to be more severe. Methods that at least partially address the specific biases that exist in such populations are discussed in a separate section of this chapter on orphanhood in populations affected by AIDS.

The method estimates the trend in mortality from data supplied by different age groups of respondent: the older the respondent, the longer ago their parents died on average. In order to convert the series of measures of survivorship obtained from different age groups into a single indicator that can be compared over time, it must be assumed that the age pattern of mortality in adulthood is represented by the chosen standard life table. To estimate the time location of these measures, it is further assumed that mortality declined linearly in terms of that standard over the period being considered. Both assumptions are likely to be problematic in populations experiencing generalized HIV epidemics.

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Before starting the analysis, one should check how many respondents stated that they did not know whether their mother or, more commonly, father was alive or failed to answer the questions at all. The response rate on these questions is usually very high and one can simply exclude from the analysis those respondents who either answered 'don't know' or did not answer the question. In effect, this amounts to assuming that the proportion of these respondents' parents that have died is the same as for respondents that answered the question. However, a few surveys have collected sufficiently incomplete data to suggest that non-response bias could be a substantial problem. For example, it is possible that most people who fail to answer the question have dead parents. If this is the case, such unreported orphans could represent a substantial proportion of all orphans, particularly in the younger age groups, producing a substantial downward bias in the final estimates of mortality.

One useful check on the quality of the orphanhood data is to compare the responses of male and female respondents of the same age. One would not expect the proportion of parents that have died to differ significantly between men and women of the same age. If the proportions diverge among older respondents, this could reflect gender differences in patterns of age misreporting or could indicate that the gender that reports fewer dead parents (usually the men) is more likely to lose touch with their families and is assuming wrongly that some parents remain alive who have died.

## CAVEATS AND WARNINGS

- The estimates derived from the orphanhood method are conditional survivorship probabilities, that is to say probabilities of surviving across an interval in adulthood conditional on being alive at the start of the interval. To obtain a complete life table, estimates of survivorship from birth to adulthood must be calculated using another source of data on child mortality.

- Deaths of parents do not occur at one point in time but may have occurred at any time between the respondents' birth and when they were interviewed. Thus, the orphanhood method can only indicate the smoothed trend in adult mortality and will fail to capture short-term mortality crises or abrupt reversals in the trend in mortality such as those resulting from AIDS after the onset of a generalized HIV epidemic.
- The most up-to-date mortality estimates that can be produced using orphanhood data are made from those on children aged 5 to 14. Typically only a few per cent of children are maternal orphans and the resulting estimates can have wide confidence intervals even in large sample surveys.
- In a number of applications in East Africa and elsewhere, the orphanhood method has yielded results that indicate implausibly rapid declines in mortality and gross inconsistencies between the estimates from successive enquiries. This appears to be due to 'the adoption effect', that is under-reporting of orphanhood among those whose parents die when they are very young (Blacker 1984; Blacker and Gapere 1988; Hill 1984; Timæus 1986). Children who are orphaned at a young age tend to be reared by other relatives and are often enumerated as their own children. This means they are enumerated as having a living parent and can give rise to very low mortality estimates. Misreporting appears to be particularly common when the mother dies. As the respondents get older, the chance that their foster, adoptive or step-parent has died, as well as their biological parent, increases. This implies that the bias is most pronounced for young children, whose substitute parent is very likely to be alive. Thus, if it is not recognized, the adoption effect leads to the underestimation of recent adult mortality and an overestimate of the extent of mortality decline.
- Although estimates can be made using data on respondents aged in their forties, the parents of many of these respondents are elderly and have very high mortality. This means that the precision with which one can estimate both mortality from parental survival data and the reference dates of the estimates is inherently much lower than it is for younger respondents.
- The orphanhood method does not involve the assumption that the population is closed to migration. Nevertheless, it can be difficult to interpret orphanhood-based estimates of adult mortality for sub-national geographic units, such as urban and rural areas or districts, or for respondents with particular socio-economic characteristics. Some or many of the respondents' parents will live in different places from the respondents themselves and their socio-economic characteristics may differ from those of their children. However, parents and children usually do share the same ethnic identity.

## APPLICATION OF METHOD

### Maternal orphanhood

#### *Step 1a: Calculate the mean age of childbearing of women, $\bar{M}^f$*

To apply the orphanhood method, one requires an estimate of the average age at which the parents had their children in order to control for variation in the age range over which they have been exposed to the risk of dying. Women's mean age of childbearing is usually calculated from census or survey data on recent births by five-year age group at interview of the women giving birth. The measure is simply the average age of women giving birth calculated without adjusting for the age structure of the population using the following formula:

$$\bar{M}^f = \frac{\sum_{x=15}^{45} {}_5B_x\,(x+2)}{\sum_{x=15}^{45} {}_5B_x}.$$

In this equation, ${}_5B_x$ represents births to women in the age group $x$ to $x+4$ completed years and $(x+2)$ represents the mid-point of the age group of women with a half-year downward shift to allow for the fact that women giving birth in the year before interview did so 6 months ago, on average, and were 6 months younger at that time. This calculation can be done in the accompanying Excel workbook (see website). If the data used to calculate $\bar{M}^f$ are tabulated by women's age at giving birth, the mid-point of each age group would become $x+2.5$.

There is no need to adjust the births data for reference-period errors before calculating $\bar{M}^f$. Moreover, the mortality estimates are not very sensitive to bias in this indicator. However, if evidence exists that the age pattern of births has been distorted severely by women exaggerating their ages, the number of births by age could be recomputed from an adjusted age distribution and adjusted fertility distribution before calculating $\bar{M}^f$.

In principle, the mean age of motherhood should refer

to the time at which the respondents were born, which may be any time between 5 and 45 years before the collection of the orphanhood data. An estimate based on fertility data collected in the same enquiry that asked about orphanhood should be adequate in populations which have yet to experience substantial fertility decline. If fertility is believed to have fallen and earlier census or survey data exist, $\bar{M}^f$ could also be calculated from the earlier data to determine if it has changed. If it has, then the best way of deciding on final values of $\bar{M}^f$ for the estimation of adult mortality will depend on what data are available and the pattern of change in fertility. One option might be to calculate $\bar{M}^f$ from data collected at about the time that fertility began to fall and use that value for age groups of respondents born then or earlier and to interpolate linearly between that value and the current one to estimate $\bar{M}^f$ for younger age groups of respondents.

***Step 2a: Calculate the conditional life table survivorship ratios for women***

The workbook (see website) contains separate sheets for the calculation of the survivorship ratios for adult women (maternal orphanhood) and adult men (paternal orphanhood). The sheets are set up both to analyse data on male and female respondents separately and to combine them to produce estimates based on the data supplied by both sexes. Either the number of respondents of each sex by five-year age group with living mothers and number answering the question or the proportions with living mothers calculated from them should be entered into the maternal orphanhood sheet. Data on respondents aged 5 to 49 years can be used.

Women's survivorship is estimated between a lower age of 25 and age $25+n$, where $n$ is the upper limit of each successive age group of respondents. The following regression equation and the coefficients shown in Table 22.1 are used:

$$ {}_{n}p_{25} = a(n) + b(n)\bar{M}^f + c(n)\,{}_{5}S_{n-5}\,. $$

where ${}_{5}S_{n-5}$ is the proportion of respondents in the age group from exact age $n-5$ to exact age $n$ with living mothers. For example, when $n$ is 10, life table survivorship is estimated over the 10-year age interval from exact age 25 to exact age 35 using data on survival of mothers supplied by respondents aged 5–9 years.

*TABLE 22.1 Coefficients for the estimation of women's survivorship from the proportions of respondents with living mothers*

| $n$ | $a(n)$ | $b(n)$ | $c(n)$ |
|---|---|---|---|
| 10 | –0.2894 | 0.00125 | 1.2559 |
| 15 | –0.1718 | 0.00222 | 1.1123 |
| 20 | –0.1513 | 0.00372 | 1.0525 |
| 25 | –0.1808 | 0.00586 | 1.0267 |
| 30 | –0.2511 | 0.00885 | 1.0219 |
| 35 | –0.3644 | 0.01287 | 1.0380 |
| 40 | –0.5181 | 0.01795 | 1.0753 |
| 45 | –0.6880 | 0.02343 | 1.1276 |
| 50 | –0.8054 | 0.02721 | 1.1678 |

Source: Timæus (1992)

**Paternal orphanhood**

***Step 1b: Calculate $\bar{M}^m$ the mean age of childbearing of men***

In principle, the best way to calculate the mean age at which men have children is to calculate the mean age of the partners of women who gave birth in a defined year. However, this method requires data on women to be linked to those on their partners, which is impossible in many data sets.

Thus, the mean age at which men have children is usually estimated by adding an index of the difference between the ages of men and women who are having children to the mean age of childbearing of women:

$$ \bar{M}^m = \bar{M}^f + d\,. $$

One estimate of this difference that can be readily calculated from census data is the difference between the median ages of currently married men and currently married women. It is more appropriate than the difference between the singulate mean ages at marriage of men and women in populations in which marital dissolution or polygynous marriage is common. The median is used rather than the mean so that differential age exaggeration by older respondents, who are probably no longer bearing children anyway, does not distort the estimate.

This approach to the estimation of the mean age of men at the birth of their children assumes that the ages of the fathers of children born to unmarried women are the same, on average, as the ages of the fathers of children born to married women. They may not be and this could introduce a significant bias into the estimate of $\bar{M}^m$ in populations in which childbearing outside marriage is common. While it is difficult to think of a solution to this problem, fortunately

the mortality estimates are not very sensitive to errors in the estimate of $\bar{M}^m$.

***Step 2b: Calculate the conditional life table survivorship ratios for men***

The Excel workbook (see website) contains separate sheets for the calculation of the survivorship ratios for adult women and adult men. Either the number of respondents by five-year age group with living fathers and number answering the question or the proportions with living fathers calculated from them should be entered into the paternal orphanhood sheet. Data are used only on respondents aged 5 to 44 years.

Each estimate of men's survivorship is produced using data on two adjacent five-year age groups. For every pair of age groups except the youngest pair analysed, men's survivorship is measured from age 35 to $35+n$, where $n$ is the midpoint of the pair of age groups, using the following regression equation and the coefficients shown in Table 22.2:

$$ {}_n p_{35} = a(n) + b(n)\bar{M}^m + c(n)\,{}_5S_{n-5} + d(n)\,{}_5S_n . $$

*TABLE 22.2 Coefficients for the estimation of men's survivorship from the proportions of respondents with living fathers*

| $n$ | $a(n)$ | $b(n)$ | $c(n)$ | $d(n)$ |
|---|---|---|---|---|
| 10 | –0.5578 | 0.00040 | 1.4708 | 0.0698 |
| 15 | –0.4013 | 0.00576 | 1.5602 | –0.3522 |
| 20 | –0.3329 | 0.01031 | 0.6656 | 0.3419 |
| 25 | –0.4726 | 0.01559 | 0.2161 | 0.7896 |
| 30 | –0.7056 | 0.02076 | 0.1997 | 0.9066 |
| 35 | –0.9153 | 0.02493 | 0.3484 | 0.8631 |
| 40 | –0.9950 | 0.02635 | 0.4269 | 0.8263 |

Source: Timæus (1992)

For example, when $n$ is 10, life table survivorship is estimated over the 10-year age interval from exact age 35 to exact age 45 using the data on survival of fathers supplied by respondents in the two age groups 5–9 years and 10–14 years.

For the youngest pair of age groups, however, this survivorship ratio is much higher than the proportions of fathers alive, which makes the estimates very sensitive to the exact age pattern of mortality in the population. More robust estimates are obtained using the equation:

$$ {}_{15}p_{35} = a(10) + b(10)\bar{M}^m + c(10)\,{}_5S_5 + d(n)\,{}_5S_{10} . $$

In other words, the same survivorship ratio, ${}_{15}q_{35}$, is estimated from data on respondents aged both 5–14 and 10–19.

**Maternal and paternal orphanhood**

***Step 3: Convert the survivorship ratios into estimates of the level of mortality***

To infer mortality trends from a series of survival ratios, ${}_np_b$, obtained from different age groups of respondents and referring to different dates, it is necessary to convert them all into a common index of mortality that can be compared over time. This is done by fitting a 1-parameter model life table to each measure and obtaining the common mortality index from the model.

A wide range of indices have been used for this purpose, including the level parameters of various systems of model life tables, survivorship ratios such as ${}_{35}p_{30}$, which correspond fairly closely to the age range of the estimates yielded by the orphanhood method, life expectancy at various ages between 5 and 30, and temporary life expectancy between ages 25 and 70, ${}_{45}e_{25}$. Using the parameters of the models has the advantage of emphasizing that the full life table is being estimated by fitting a model, rather than measured directly. The measures of life expectancy summarize survivorship across adulthood as a whole, while using survivorship ratios or temporary life expectancies avoids extrapolation into old age from measures for younger adults. Increasingly, in recent years, the estimates have been presented in terms of the probability of dying between exact ages 15 and 60, ${}_{45}q_{15}$, as this measure has found favour with several international agencies as a summary indicator of the mortality of young and middle-aged adults.

In the applications of the orphanhood method presented here the survivorship ratios are converted into the $\alpha$ parameter of a 1-parameter system of relational logit model life tables, and then into either estimates of the probability of dying between exact ages 15 and 60, ${}_{45}q_{15}$, or the probability of dying between exact ages 30 and 60, ${}_{30}q_{30}$. The latter measure is useful for comparing series of orphanhood estimates with each other without extrapolating from mortality across one age range to mortality across another by more than is necessary. The parameters of the 1-parameter models are calculated from the estimates of ${}_np_b$ as

$$ \alpha = -\frac{1}{2}\ln\left(1 + \frac{{}_np_b - \dfrac{1}{l^s_{b+n}} - \dfrac{1}{l^s_b}}{1 - {}_np_b}\right) $$

where the estimates of ${}_np_b$ come from Step 2 with $b=25$ for the estimates of women's survivorship and $b=35$ for

those of men's survivorship and the $l_x^s$ values come from a standard life table. Thus, one obtains a series of values of $\alpha$ corresponding to the measures of conditional survivorship made from data on the different age groups of respondents. Higher values of $\alpha$ correspond to higher mortality. Then, for each $\alpha$,

$$ {}_{45}q_{15} = 1 - \frac{1 + e^{2(\alpha + Y^s(15))}}{1 + e^{2(\alpha + Y^s(60))}} $$

or

$$ {}_{30}q_{30} = 1 - \frac{1 + e^{2(\alpha + Y^s(30))}}{1 + e^{2(\alpha + Y^s(60))}}. $$

The workbook (see website) can calculate $\alpha$, ${}_{45}q_{15}$, and ${}_{30}q_{30}$ using either a standard from the General family of United Nations model life tables or one from any of the four families of Princeton model life tables. The standard life table should be chosen to have an age pattern of mortality within adulthood that resembles that of the population being studied. Another life table can be used as a standard if there is reason to believe that it resembles more closely the pattern of adult mortality in the population being studied. The most suitable life table may not be from the family of models that best captures the relationship between child and adult mortality. If nothing is known about the age pattern of mortality in adulthood, use of the United Nations General or Princeton West models is recommended.

***Step 4: Calculate the time location of the estimates***

Each survivorship ratio produced by the orphanhood method represents an average of the mortality prevailing during the $N$ years that parents have been exposed to the risk of dying. The point of time at which the ratio equalled period survivorship depends on the level of mortality and can be estimated from the proportion of the mothers or fathers that are alive, the age of the respondents, and the mean age at childbearing of the parents.

The only additional information required to calculate the dates to which each of the adult morality estimates refer is the date on which the inquiry that asked about the survival of mothers and fathers was conducted. This can be calculated as the average of the dates on which the interviews took place or taken as the mid-point of the period of fieldwork if exact dates of interview are not available.

The estimates are calculated as the date of fieldwork minus $T$, where $T$ is calculated for women as

$$ T = \frac{N}{2}\left(1 - \frac{1}{3}\ln\left({}_5S_{n-5}\right) + \frac{1}{3}\ln\left(\frac{80 - \bar{M}^f - N}{80 - \bar{M}^f}\right)\right). $$

Because the mortality of the mothers is estimated from data on a single five-year age group of respondents, their average length of exposure to the risk of dying, $N$, is the mid-point of that age group (i.e. 7.5 years, 12.5 years, and so on). However, because pairs of age groups are used to estimate men's mortality, $N$ becomes the age dividing the two age groups (i.e. 10 years, 15 years, and so on). Moreover, because fathers might have died during the 9 months before the respondents were born, $T$ becomes

$$ T = \frac{(N + 0.75)}{2}\left(1 - \frac{1}{3}\ln\left(\sqrt{{}_5S_{n-5} \cdot {}_5S_n}\right) + \frac{1}{3}\ln\left(\frac{80 - \bar{M}^m - N}{80 - (\bar{M}^m - 0.75)}\right)\right). $$

For both men's and women's mortality, the time location of the estimates obtained from orphanhood data on respondents in the different age groups after age 30 is usually similar. Although the date at which the parents could first have died gets earlier as the respondents' ages rise, this is cancelled out by an increase in the proportion of deaths that occurred shortly before the survey as the parents move into old age. Sometimes the estimated time location of mortality indicators obtained from data on the oldest age groups is more recent than the time location of the mortality indicators obtained from data on respondents who are five years younger. Given the approximate nature of the mapping of cohort measures of mortality into period estimates, it is often advisable to ignore the estimates from the oldest age groups or to average the final two or three estimates and the dates to which they refer to produce a single measure.

## WORKED EXAMPLE

This example uses data on the survival of mothers and fathers collected in the 1997 Census of Iraq, available from IPUMS. Only the results for respondents of both sexes together are shown in the tables.

***Step 1a: Calculate the mean age of childbearing of women,*** $\bar{M}^f$

The mean age of childbearing is a straightforward average of the ages of women giving birth and can either be calculated as such from individual-level data or estimated approximately from a tabulation of births by five-year age group of mother (Table 22.3).

$$\bar{M}^f = \frac{17338360}{613080} = 28.28\,.$$

***Step 2a: Calculate the conditional life table survivorship ratios of women***

These survivorship ratios are shown in the sixth column of Table 22.4 and are calculated from the proportions in the fourth column using the regression coefficients shown in Table 22.1 and the estimate of $\bar{M}^f$ of 28.28 from Step 1a. For example, for respondents aged 5–9,

$$_{10}p_{25} = -0.2894 + 0.00125 \times 28.28 + 1.2559 \times 0.9916 = 0.9913\,.$$

Note that each life table measure is similar in value to the proportion from which it was calculated.

***Step 1b: Calculate the mean age of childbearing of men,*** $\bar{M}^m$

The mean age of childbearing of men is calculated by adding the difference between the median ages of currently married men and women to the mean age of childbearing of women. It can be seen from Table 22.5 that the median age of currently married men falls between the mid-point of the age group 30–34 and the mid-point of the age group 35–39. By linear interpolation,

Median for men

$$= 32.5 + 5 \times \frac{0.5 - 0.4105}{0.5208 - 0.4105} = 36.56$$

and

Median for women

$$= 27.5 + 5 \times \frac{0.5 - 0.3780}{0.5372 - 0.3780} = 31.33\,.$$

*TABLE 22.3 Calculation of the mean age at childbearing, Iraq Census, 1997*

| Age group | Births in the last year $B(i)$ | Mid-point age $N$ | $B(i)*N$ |
|---|---|---|---|
| 15–19 | 50,720 | 17 | 862,240 |
| 20–24 | 147,040 | 22 | 3,234,880 |
| 25–29 | 174,350 | 27 | 4,707,450 |
| 30–34 | 129,860 | 32 | 4,155,520 |
| 35–39 | 66,440 | 37 | 2,458,280 |
| 40–44 | 35,900 | 42 | 1,507,800 |
| 45–49 | 8,770 | 47 | 412,190 |
| **TOTALS** | **613,080** | | **17,338,360** |

*TABLE 22.4 Estimation of women's survivorship, the dates at which these estimates are located in time, and corresponding estimates of* $\alpha$ *and* $_{30}q_{30}$ *from maternal orphanhood, Iraq, 1997 Census, both sexes*

| Age group | Total women | Mother alive | Proportion alive | Age $n$ | $\frac{l(25+n)}{l(25)}$ | Level ($\alpha$) | Probability of dying ($_{30}q_{30}$) | Date |
|---|---|---|---|---|---|---|---|---|
| 5–9 | 2,834,210 | 2,810,420 | 0.9916 | 10 | 0.9913 | –0.756 | 0.072 | 1994.2 |
| 10–14 | 2,391,580 | 2,347,190 | 0.9814 | 15 | 0.9826 | –0.670 | 0.084 | 1992.1 |
| 15–19 | 2,181,840 | 2,108,280 | 0.9663 | 20 | 0.9709 | –0.625 | 0.090 | 1990.1 |
| 20–24 | 1,827,030 | 1,719,880 | 0.9414 | 25 | 0.9514 | –0.561 | 0.101 | 1988.5 |
| 25–29 | 1,615,160 | 1,455,150 | 0.9009 | 30 | 0.9198 | –0.497 | 0.112 | 1987.0 |
| 30–34 | 1,235,120 | 1,035,480 | 0.8384 | 35 | 0.8698 | –0.438 | 0.124 | 1985.9 |
| 35–39 | 788,170 | 601,590 | 0.7633 | 40 | 0.8103 | –0.447 | 0.122 | 1985.4 |
| 40–44 | 817,460 | 519,580 | 0.6356 | 45 | 0.6910 | –0.375 | 0.137 | 1985.5 |
| 45–49 | 603,840 | 298,760 | 0.4948 | 50 | 0.5419 | –0.346 | 0.144 | 1988.3 |

***Step 2b: Calculate the conditional life table survivorship ratios of men***

These survivorship ratios are shown in the sixth column of Table 22.6 and are calculated from the proportions in the fourth column using the regression coefficients shown in Table 22.2 and the estimate of $\bar{M}^m$ of 33.51 from Step 1b. For example, for the final estimate in Table 22.6,

*TABLE 22.5 Ages of currently married men and women, Iraq, 1997 Census*

| Age group | Married men | Married women | Cumulative proportion of men | Cumulative proportion of women |
|---|---|---|---|---|
| 10–14 | 1,380 | 7,300 | 0.0005 | 0.0024 |
| 15–19 | 34,140 | 171,320 | 0.0122 | 0.0583 |
| 20–24 | 192,160 | 429,420 | 0.0782 | 0.1984 |
| 25–29 | 471,780 | 550,700 | 0.2401 | 0.3780 |
| 30–34 | 496,440 | 488,090 | 0.4105 | 0.5372 |
| 35–39 | 321,400 | 345,680 | 0.5208 | 0.6500 |
| 40–44 | 375,650 | 336,270 | 0.6497 | 0.7597 |
| 45–49 | 279,350 | 245,010 | 0.7456 | 0.8396 |
| 50–54 | 224,610 | 164,470 | 0.8227 | 0.8933 |
| 55–59 | 166,800 | 115,380 | 0.8800 | 0.9309 |
| 60–64 | 105,760 | 75,470 | 0.9163 | 0.9555 |
| 65–69 | 99,360 | 63,890 | 0.9504 | 0.9764 |
| 70–74 | 59,000 | 37,220 | 0.9706 | 0.9885 |
| 75–79 | 41,110 | 19,700 | 0.9847 | 0.9949 |
| 80+ | 44,430 | 15,570 | 1.0000 | 1.0000 |
| **TOTAL** | **2,913,370** | **3,065,490** | | |

$$ {}_{40}p_{35} = -0.9950 + 0.02635 \times 33.51 + 0.4269 \times 0.4818 + 0.8263 \times 0.3432 = 0.3772 . $$

Both the estimate of 0.9336 (made from the proportions 0.9622 and 0.8999) and the estimate of 0.9000 (made from the proportions 0.8999 and 0.8394) are measures of ${}_{15}p_{35}$, whereas the following estimate, 0.8398, measures ${}_{20}p_{35}$, and so on.

***Step 3: Convert the survivorship ratios into estimates of the level of mortality***

The seventh columns of Tables 22.4 and 22.6 contain $\alpha$, the level parameter of the system of relational logit model life tables with a General United Nations model life table for both sexes with $e_0 = 60$ as its standard. For example, $\alpha$ is calculated from the estimate of ${}_{10}p_{25}$ for women as

$$ \alpha = -\frac{1}{2}\ln\left(1 + \frac{\frac{0.9913}{0.8335} - \frac{1}{0.86313}}{1 - 0.9913}\right) = -0.756 . $$

The eighth columns of Tables 22.4 and 22.6 contain ${}_{30}q_{30}$, the probability of dying between exact ages 30 and 60. Having calculated $\alpha$ from ${}_{10}p_{25}$ to be –0.756, the corresponding measure of ${}_{30}q_{30}$ is

$$ {}_{30}q_{30} = 1 - \frac{1 + e^{2(-0.756 - 0.8655)}}{1 + e^{2(-0.756 - 0.3064)}} = 0.072 . $$

*TABLE 22.6 Estimation of men's survivorship, the dates at which these estimates are located in time, and corresponding estimates of $\alpha$ and ${}_{30}q_{30}$ from paternal orphanhood, Iraq, 1997 Census, both sexes*

| Age group | Total women | Father alive | Proportion alive | Age $n$ | $\frac{l(35+n)}{l(35)}$ | Level ($\alpha$) | Probability of dying (${}_{30}q_{30}$) | Date |
|---|---|---|---|---|---|---|---|---|
| 5–9 | 2,834,210 | 2,727,110 | 0.9622 | 15 | 0.9336 | –0.216 | 0.175 | 1992.7 |
| 10–14 | 2,391,580 | 2,152,090 | 0.8999 | 15 | 0.9000 | 0.045 | 0.251 | 1990.6 |
| 15–19 | 2,181,840 | 1,831,390 | 0.8394 | 20 | 0.8398 | 0.057 | 0.254 | 1988.7 |
| 20–24 | 1,827,030 | 1,435,320 | 0.7856 | 25 | 0.7707 | 0.026 | 0.245 | 1987.0 |
| 25–29 | 1,615,160 | 1,127,370 | 0.6980 | 30 | 0.6629 | 0.043 | 0.250 | 1985.5 |
| 30–34 | 1,235,120 | 726,870 | 0.5885 | 35 | 0.5409 | 0.025 | 0.244 | 1984.6 |
| 35–39 | 788,170 | 379,770 | 0.4818 | 40 | 0.3772 | 0.065 | 0.257 | 1984.8 |
| 40–44 | 817,460 | 280,540 | 0.3432 | | | | | |

***Step 4: Calculate the time location of the estimates***

The dates in the ninth column of Tables 22.4 and 22.6 are calculated by subtracting the time location of the estimates from the date of the 1997 Census of Iraq, which was taken on 16 October. As a decimal, this becomes 1997.79.

The time location of each estimate is calculated from the respondents' ages, the appropriate estimate of $\bar{M}$, and the proportions surviving. For mothers, exposure starts at the mean age of childbearing and extends from that age to the average age of the respondents supplying the data from which the survivorship ratio, *N,* was calculated. The initial age at exposure of fathers is $\bar{M}^m - 0.75$ but their duration of exposure is $N + 0.75$ years. For example, for the first estimate of men's survivorship, the calculation is

$$T = \frac{(10+0.75)}{2}\left(1-\frac{1}{3}\ln\left(\sqrt{0.9622\cdot 0.8999}\right)\right.$$
$$\left.+\frac{1}{3}\ln\left(\frac{80-33.51-10}{80-(33.51-0.75)}\right)\right)$$
$$=\frac{10.75}{2}(1-0.0621)=5.04\ .$$

Thus, the date to which this first estimate refers is $1997.79 - 5.04 = 1992.75$.

The time location of the estimate of women's mortality calculated from data on respondents aged 45–49 is nearly three years later than the time location of the estimate calculated from data on respondents aged 40–44. As in many applications of the orphanhood method, the data on this final age group should be discarded.

## DIAGNOSTICS, ANALYSIS AND INTERPRETATION

### Checks and validation

The number of respondents who stated that they did not know whether their mother or father is alive or who did not answer the questions at all should be checked before they are dropped from the analysis. If many of the respondents failed to respond to these questions, the data supplied by those respondents who did answer them may not be representative of the population as a whole. Moreover, a high level of non-response may indicate that either the field staff or the respondents experienced difficulty with the questions. Thus, even when answers were supplied they may be rather unreliable. If a high level of item non-response exists, it can be illuminating to determine whether it is concentrated among a minority of field staff or a certain type of respondent.

If information about the survival of mothers and fathers has been collected from both male and female *respondents* in a census or a large-scale survey with small sampling errors, it is possible to tabulate the proportions of mothers and fathers alive separately for respondents of each sex in order to compare the consistency of their reports. While consistency of reporting does not guarantee accuracy, statistically significant differences between the proportions obtained from male or female respondents imply that at least one of the sexes, and possibly both of them, is answering the questions inaccurately.

It is fairly common to find that women report lower proportions of living parents than men. Some analysts believe that this is because women stay in closer contact with their parents than men and that some men are stating that their parents are alive because they do not know that they have in fact died. If correct, this would imply that the data supplied by women are more accurate. However, no strong evidence exists to support this interpretation and other errors, notably differential age misreporting by male and female respondents, may also produce inconsistencies between the proportions reported by men and women. In surveys of a few thousand households, moreover, the sampling errors of the estimates for the first few age groups may be quite large relative to the proportion of mothers or fathers that have died. Thus, if no clear reason exists to focus on the data supplied by female respondents, it can be advisable to base the final estimates of adult mortality on the combined responses of both sexes.

### Interpretation

The results of the analysis of the orphanhood data from the 1997 Census of Iraq are portrayed graphically in Figure 22.1. The craft of interpreting such plots is not just to read off the putative trend, but to decide which features reflect changes in mortality and which reflect problems with the data or the method.

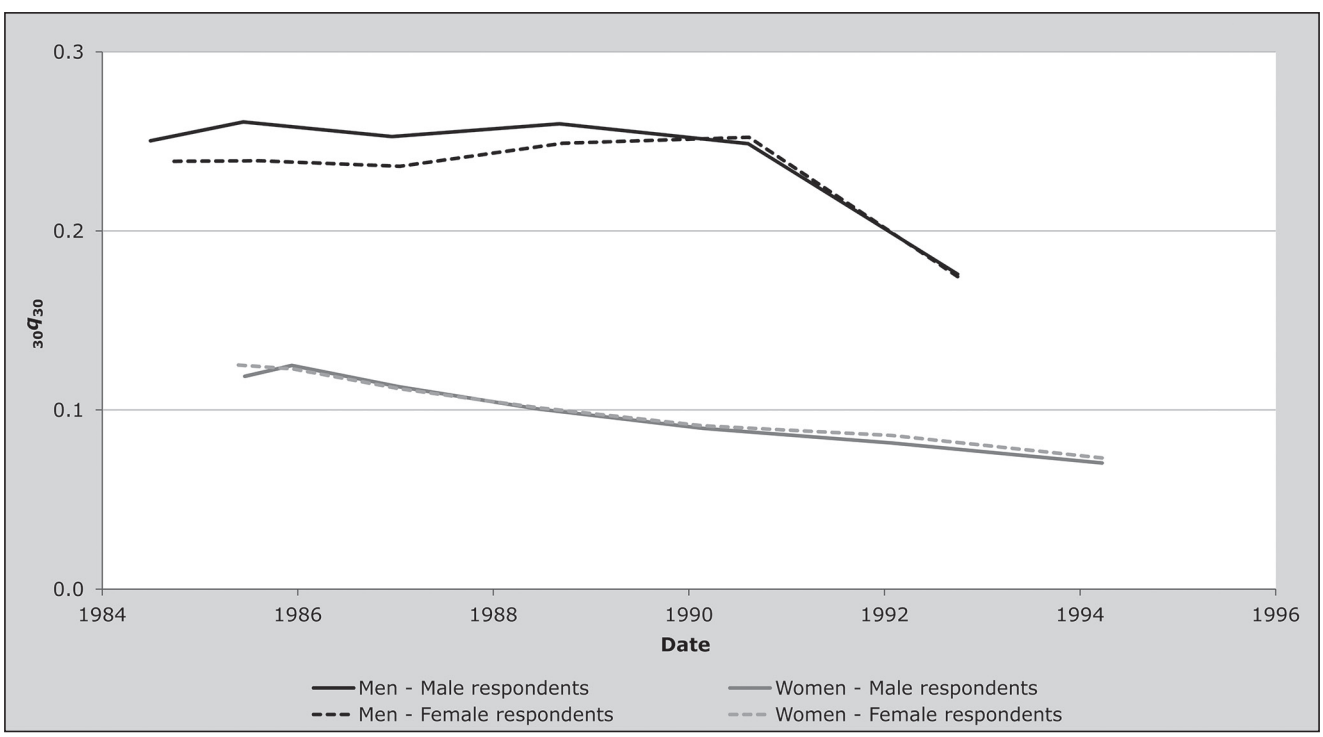
The estimates made from the data supplied by male and female respondents in this application of the method are similar, which provides some reassurance as to the quality of both age reporting and reporting on the survival of parents in Iraq. Nevertheless, adult women are slightly more likely to report that their father is alive than adult men of the same age and the size of this inconsistency in the reports

widens as the respondents' ages increase. This discrepancy is the opposite of the pattern observed in most populations. It could be explained by a stronger tendency for women than men to overstate their age or a stronger tendency for men than women to understate theirs, but both patterns of age misreporting seem unlikely. Alternatively, either some adult men could have reported their living fathers as dead or some adult women could have reported their dead fathers as alive. While this last possibility is perhaps the most plausible of the four explanations, there is no strong evidence for it. Perhaps the best course of action is to assume that both men's and women's reports on their fathers are slightly biased in opposite directions and use the estimates calculated from data on respondents of both sexes combined.

Whichever set of results one adopts, the orphanhood estimates suggest that the mortality of adult women in Iraq underwent a steady decline between the mid-1980s and mid-1990s and was only slightly higher than in most Western countries by the latter date. The smoothness of the trend is reassuring but the low level of mortality is surprising. Without other evidence to corroborate these results, one should be cautious about accepting them at face value as they could be biased downward by failure to report the deaths of some dead mothers, particularly those who died when the respondents were young children. This tends to produce an exaggerated impression of the extent to which mortality has declined.

In contrast, the mortality of adult Iraqi men in this period is estimated to be much higher than that of women. Moreover, the results suggest that men's mortality stagnated through much of the 1980s before falling quite sharply in the early 1990s. It is likely that this pattern reflects the impact on orphanhood of the casualties in the 1980–1988 Iran–Iraq war, in which a quarter to half a million Iraqi soldiers are believed to have died. Whereas vital statistics data usually distinguish between civilian and military deaths, in census and survey data both appear together. Thus, the most recent estimate of men's mortality may be biased down by underreporting of dead fathers, but could be much lower than the others because nearly half the respondents on which it is based were born after the end of the war. Their fathers, therefore, must have survived it. In contrast, all the

***FIGURE 22.1*** *Trends in the probability of dying between exact ages 30 and 60 estimated from orphanhood, Iraq, 1997 Census*



other estimates are based on the reports of respondents born before 1988. The fathers of the respondents aged 30–39, who supplied the information used to produce the earliest estimate, were about 50 years old on average at the outbreak of the war. Most of them would have been too old to have served at the front.

It should be recalled, however, that the orphanhood method tends to smooth out fluctuations in mortality because each estimate is based on deaths occurring over the entire period between the birth of the respondents and the time that the data were collected. In addition, the estimated trend can be distorted somewhat if the model life table adopted for the calculation of ${}_{30}q_{30}$ is inappropriate. Thus, these results cannot be interpreted as portraying the precise trend in men's mortality in Iraq between the early 1980s and early 1990s. For example, men's mortality may have risen somewhat before it fell or may have fluctuated appreciably from year to year during the war, depending on the number of casualties.

## DETAILED DESCRIPTION OF METHOD

### Introduction

The potential of information on orphanhood for the measurement of adult mortality was first explored by Henry (1960). If survivorship could be treated as a linear function of age, the proportion orphaned at any age $x$, $O(x)$, would to a close approximation depend on the probability of surviving from the mean age ($\bar{M}$) of parents at the birth of their children to age $\bar{M}+x$. However, this relationship underestimates the survivorship ratio (${}_x p_{\bar{M}}$) because the rapid increase in death rates with age means that survivorship curve in adulthood is convex. The lower risk of dying of those who bear children at young ages fails to compensate fully for the higher risk of dying of parents who bear children relatively late in life. To allow for this, Henry proposed an adjustment based on the variance in ages at childbearing.

The subsequent derivation of simple, robust methods for estimating mortality from orphanhood is largely associated with Brass. Hill and Blacker also played an important role in deriving ways of estimating men's mortality from paternal orphanhood. Development of the procedure occurred in a number of stages; several variants were circulated informally and eventually published in Brass (1975). The final version of Brass's implementation of the method, however, is that published in Brass and Hill (1973). It calculates measures of life table survivorship as a weighted average of the proportions of respondents with living parents in two adjoining age groups with the weights being determined by, and adjusting for, the mean age of childbearing ($\bar{M}$) in the population.

Subsequently several researchers proposed regression methods for predicting women's mortality from the proportions of respondents with surviving mothers (Hill and Trussell 1977; Palloni and Heligman 1985). Timæus (1992) demonstrated that the weighting factors tend to underestimate mortality when it is higher than in the standard used to derive them and to overestimate mortality when it is lower than in the standard. A regression-based approach should usually produce more accurate estimates, particularly of men's mortality, because the intercept term provides additional flexibility with which to model more accurately the relationship between parental survival and life table survivorship at differing levels of mortality. Therefore, Timæus (1992) proposed regression coefficients for men's mortality, together with coefficients derived using consistent assumptions for the estimation of women's mortality. It is this variant of the orphanhood method that is presented here.

### Mathematical exposition – orphanhood method

Let the number of children born $a$ years before a demographic enquiry to women aged $y$ when they gave birth be ${}_aV_y$. The probability of the children being alive at the time of the enquiry is $l(a)$ and the probability that their mothers are still alive is $l(y+a)/l(y)$. Assuming that the mortality of orphans and children with living parents is the same, the proportion of respondents aged a with living mothers, $S(a)$, is

$$S(a) = \frac{l(a)\int_s^\omega {}_aV_y\, l(y+a)/l(y)\,dy}{l(a)\int_s^\omega {}_aV_y\,dy} \qquad (1)$$

where integration is over all ages at child bearing $s$ to $\omega$.

The number of children born to women aged $y$ is a function of the number of women aged $y$ and the fertility rate at age $y$. If we assume a stable age structure, then ${}_aV_y = B(0)e^{-r(y+a)}l(y)f(y)$, where $r$ is the rate of natural increase and $f(y)$ the mothers' fertility rate at age $y$. Therefore, the proportion of respondents in a five-year age group $x$ to $x+5$ with living mothers is

$$ {}_5S_x = \frac{\int_x^{x+5} e^{-ra} l(a) \int_s^{\omega} e^{-ry} f(y) l(y+a) dy da}{\int_x^{x+5} e^{-ra} l(a) \int_s^{\omega} e^{-ry} f(y) l(y) dy da}. $$

Fathers' exposure extends from the conception of the respondents to their current age. Making allowance for the possibility that a father may die between the conception and birth of his child gives a slightly more complicated equation for the proportion of respondents in a five-year age group $x$ to $x+5$ with living fathers:

$$ {}_5S_x = \frac{\int_x^{x+5} e^{-ra} l(a) \int_s^{\omega} e^{-ry} f(y) l(y) l(y+a)/l(y-0.75) dy da}{\int_x^{x+5} e^{-ra} l(a) \int_s^{\omega} e^{-ry} f(y) l(y) dy da}. $$

In this equation, $f(y)$ represents the age-specific fertility schedule, and $l(a)$ the life table survivorship, of men rather than of women and the ages between which childbearing occurs $s$ and $\omega$ are also those of men. There is no straightforward way of integrating either the numerator or denominator of these expressions for ${}_5S_x$ and they have to be evaluated numerically.

**Implementation of the method**

In principle, there is no reason why Equation 1 should not be solved directly for life table survivorship using the Excel Solver routine or a similar tool and an age structure for those giving birth, ${}_aV_y$ that is appropriate for the population under study. To arrive at a unique solution, an assumption still has to be made about the age pattern of mortality within adulthood such as which standard to adopt in a 1-parameter system of relational model life tables. In practice, estimates are usually produced using a regression model that controls for $\bar{M}$ which has been fitted to simulated data on parental survival generated for populations with a wide range of age structures and fertility and mortality schedules (Timæus 1992).

The proportions of mothers alive in age groups $x$ to $x+5$ are more closely correlated with women's survivorship over the age range 25 to $x+5$ than any other survivorship ratio and so the latter measure can be estimated from the proportions more accurately than any other:

$$ {}_np_{25} = a(n) + b(n)\bar{M}^f + c(n){}_5S_{n-5}. $$

Particularly in populations with fairly low fertility and high mean ages of childbearing, the precision of the estimates for men is increased significantly by estimating survivorship from the proportions of fathers alive in two adjacent age groups. The additional age group captures the effects of variations in the rate at which mortality rises with age. At most ages the proportions of fathers alive in age groups $x$ to $x+10$ are more closely correlated with men's survivorship over the age range 35 to $x+5$ than any other survivorship ratio. Thus, the regression equation is

$$ {}_np_{35} = a(n) + b(n)\bar{M}^m + c(n){}_5S_{n-5} + d(n){}_5S_n. $$

The first pair of age groups, those aged 5–9 and 10–14, predict survivorship from 35 to 50 more precisely than survivorship from 35 to 45. Thus the equation becomes

$$ {}_{15}p_{35} = a(10) + b(10)\bar{M}^m + c(10){}_5S_5 + d(n){}_5S_{10}. $$

**Mathematical exposition – time location of the estimates**

Equation 1 can be rewritten as a weighted average of the cohort survivorship ratios:

$$ S(a) = \int_s^{\varpi} {}_av_y \, {}_ap_y^c dy $$

where

$$ {}_av_y = \frac{{}_aV_y}{\int_s^{\omega} {}_aV_y dy}. $$

Thus, ${}_av_y$ represents the contribution made to $S(a)$, which is the proportion of adult relatives surviving for $a$ years, by relatives who become exposed to the risk of dying at age $y$. Time location methods aim to estimate the time $T$ at which the cohort measures of survival that produced the proportion of relatives surviving, ${}_ap_y^c$, equalled the equivalent period measures, ${}_ap_y(T)$, so that

$$ S(a) = \int_s^{\varpi} {}_av_y \, {}_ap_y(T) dy. $$

If we denote the mean time since death of those dying between $y$ and $y+a$ by ${}_ag_y$, by definition

$$ {}_ag_y = \frac{\int_y^{y+a} (z - y - a)\mu(z)l(z)dz}{\int_y^{y+a} \mu(z)l(z)dz} \qquad (2) $$

where $\mu(z)l(z)$ is the life table deaths at age $z$. Brass and Bamgboye (1981) show that, if mortality schedules conform to a system of 1-parameter relational logit model life tables (Brass 1971) and if the trend in adult mortality is assumed to be linear in $\alpha$, the parameter of that system of models, then the time at which the cohort survivorship of adults equals period survivorship is a weighted average of the times since death of the respondents' relatives.

$$T = \frac{\int_{x}^{x+5} l(a) \int_{s}^{\omega} {}_{a}v_{y}\, l(y+a,T)\left(1 - {}_{a}p_{y}(T)\right) {}_{a}g_{y}\, dy\, da}{\int_{x}^{x+5} l(a) \int_{s}^{\omega} {}_{a}v_{y}\, l(y+a,T)\left(1 - {}_{a}p_{y}(T)\right) dy\, da}. \quad (3)$$

This location in time depends on the level of mortality and the ages of the relatives, but is independent of the rate of change in $\alpha$. Although Brass and Bamgboye's derivation of Equation 3 takes advantage of a relationship between changes in mortality with age and with time that is specific to a relational logit system of life tables, it possible to arrive at similar formulae for $T$ on the basis of other reasonable assumptions about the trend in mortality with time by age (Palloni, Massagli and Marcotte 1984).

Equation 3 can be evaluated numerically, using values for ${}_{a}v_{y}$ and for the life table measures chosen on the basis of observed data. To develop a straightforward procedure for estimating $T$ from observed characteristics of a population, a much simpler relationship than that described by Equation 3 must be assumed. Brass and Bamgboye (1981) argue that the change in $T$ with $a$ over limited age ranges are sufficiently close to linear that all respondents in a five-year age group can be treated as of the central age $N$. Second, they argue, at the ages and levels of mortality at which indirect methods are used to estimate adult mortality, the force of mortality increases approximately exponentially with age. As a consequence, for such applications, variation in ${}_{a}g_{y}$ with $y$ is slight. Therefore, the weighting factors for ${}_{a}g_{y}$ in Equation 2 have little effect and all adult relatives can be treated as entering exposure at their mean age of entry, $M$. To a satisfactory approximation,

$$T = {}_{N}g_{M}.$$

If survivorship in adulthood fell linearly with age, so that the same number of deaths occurred at every age, then ${}_{N}g_{M}$ would be $N/2$ whatever the value of $M$. In less extreme life tables, mortality rises with age more rapidly than this and the deaths of the relatives are concentrated at older ages and, therefore, in the recent portion of the $N$-year period. This means that the time location of the estimates is closer to the survey date than $N/2$. By substituting $e^{kz}\mu(y)l(y)$ for $\mu(z)l(z)$ in Equation 2, and expanding the right-hand side in powers of $N$, Brass and Bamgboye (1981) demonstrate that the appropriate adjustment to ${}_{N}g_{M}$ is a function of the level of mortality as measured by $k$ and the age of the respondents as measured by $N$:

$${}_{N}g_{M} \approx \frac{N}{2}\left(1 - \frac{kN}{6}\right). \quad (4)$$

Brass and Bamgboye (1981) also demonstrate that the assumption that mortality increases exponentially with age implies that, in a relational logit life table system

$$\frac{e^{kN}}{e^{k_s N}} = \frac{\left({}_{N}p_{M}\right)^{2}}{\left(l_{s}(M+N)/l_{s}(M)\right)^{2}}.$$

Solving for $kN$ and substituting this expression into Equation 4 yields an estimate of ${}_{N}g_{M}$, and therefore of $T$, of

$$T = \frac{N}{2}\left(1 - \frac{\ln\left({}_{N}p_{M}\right)}{3} + \frac{1}{3}\ln\left(\frac{l_{s}(M+N)}{l_{s}(M)}\right) - \frac{k_{s}N}{6}\right). \quad (5)$$

Thus, in this formulation, the time references of measures of conditional survivorship obtained from data on adult relatives are estimated as half the duration of exposure, $N$, reduced by a factor that depends on the level of conditional survivorship relative to a standard life table.

Having arrived at this expression for $T$ on theoretical grounds, Brass (1985) approximates ${}_{N}p_{M}$ by ${}_{5}S_{x}$ and adopts as his standard life table one in which $l_{s}(x)$ is linear over the adult ages and is taken as $(1 - x/80)/2$. As $l_{s}(x)$ is linear, $T = \frac{1}{2}N$ and $k_{s}$ becomes 0. Thus, $T$ is estimated from observed data using

$$T = \frac{N}{2}\left(1 - \frac{\ln\left({}_{5}S_{x}\right)}{3} + \frac{1}{3}\ln\left(\frac{80 - M - N}{80 - M}\right)\right). \quad (6)$$

The agreement between the estimates of $T$ produced by this simplified procedure with those obtained by evaluating Equation 2 is quite close until the age of the relatives, $M+N$,

reaches 75 years. However, as $M+N$ increases, the age pattern of mortality begins to have an important effect on the mean time since the deaths of the adult relatives, ${}_{N}g_{M}$. For example, if the force of mortality increases relatively slowly with age compared with the standard, a higher proportion of deceased relatives will have died a relatively long time ago. If there is evidence of such an age pattern of mortality or its opposite, this can be allowed for by defining a more appropriate standard by changing $\beta$, the second parameter in a relational logit system of life tables (Brass 1971). If this is done, $l_s(x)$ is no longer linear with age in the high-mortality life table used as the standard and the term in $k_s$ in Equation 5 needs to be subtracted from the estimates of $T$ obtained from Equation 6. Making the same assumptions about the form of $l(x)$ as before, the adjustment is

$$\delta T = -\frac{N}{12}(\beta - 1)\ln\left(\frac{(80-M)(80+M+N)}{(80+M)(80-M-N)}\right).$$

When using Equation 6 to estimate the time location of estimates of women's mortality from data on the survival of mother, ${}_5S_x$ becomes ${}_5S_{n-5}$, the proportion of respondents with living mothers; $N$ is the midpoint of this age group, $n-2.5$; and $M = \bar{M}$ is the women's mean age at childbearing. When calculating the time location of estimates of men's mortality from data on the survival of fathers, ${}_5S_x$ is the proportion of respondents with living fathers; $M$ becomes the fathers' mean age at the conception of their children; and $N = n + 0.75$ where $n$ is the age dividing the two age groups, ${}_5S_{n-5}$ and ${}_5S_n$, that contribute to the estimate of men's survivorship. In order to centre ${}_5S_x$ on age $n$, it is calculated as

$${}_5S_x = \sqrt{{}_5S_{n-5} \cdot {}_5S_n}.$$

## EXTENSIONS OF THE METHOD

### Orphanhood in populations affected by AIDS

Adult mortality estimates made by means of the orphanhood method in populations experiencing a generalized HIV epidemic are vulnerable to severe bias. First, HIV-positive women have lower fertility than uninfected women for both behavioural and biological reasons. Second, HIV-positive women can transmit the virus to their children during pregnancy, at delivery, or when breastfeeding. These two effects mean that HIV-infected women, who are at high risk of dying, tend to have fewer living children than the population in general. Thus, the reported proportion of mothers remaining alive is exaggerated. Because women are likely to have been infected by or to infect their husbands, orphanhood-based estimates of adult men's mortality in populations affected by HIV will also be biased downward, though to a lesser extent than those for women. A third major HIV-related bias in orphanhood estimates is that the coefficients used to convert data on parental survival into conventional life table indices of mortality assume a very different age pattern of mortality in adulthood from that found in populations experiencing an AIDS epidemic.

#### *Adjustment for HIV-related selection bias*

If estimates are available of the pertinent characteristics of the HIV epidemic in a population at the time that the respondents were born, it is possible to allow for the first two biases by adjusting the reported proportions of mothers that are alive downward (Timæus and Nunn 1997). The adjustment procedure involves the assumption that most individuals who are already infected with HIV when their children are born will die within a few years. This was a reasonable assumption to make before the recent spread of large-scale treatment initiatives in the developing world but quickly ceases to be tenable after their launch. Indeed, if such programmes were able to reduce the mortality of all HIV-positive adults to a level close to that of the general population, no adjustment would be necessary.

To adjust data collected in the 2010 round of censuses or previously, one requires estimates of the prevalence of HIV infection among women attending antenatal clinics ($P$), the vertical transmission rate of HIV ($h$), which is to say the proportion of infants who acquire HIV from their mothers, and the relative level of fertility among HIV-positive women compared with HIV-negative women ($F$). Using these measures, Timæus and Nunn (1997) derive a correction factor for adjusting the reported proportions of respondents with living mothers (${}_5S_x^*$) to obtain the proportions that would be reported if HIV-positive women had as many living children as other women (${}_5S_x$):

$${}_5S_x = \frac{1-hP}{1+\frac{1-F}{F}P}\,{}_5S_x^*.$$

The correction factor is based on the idea that the bias in the reported proportion of respondents with living mothers increases with HIV prevalence at a rate determined by $F$ and

*h*. As one would expect, the higher vertical transmission, and the lower the fertility of infected women, the greater the bias in reports on the survival of mothers. Although *F*, *P* and perhaps *h* vary with the mother's age, comparison of the results of this crude adjustment with those obtained by age-specific calculations suggests that it performs adequately (Timæus and Nunn 1997).

If few or no HIV infected people receive treatment with antiretroviral drugs, as was the case in 20th century African populations, the vertical transmission rate is typically 25 to 40 per cent, averaging about a third (De Cock, Fowler, Mercier *et al*. 2000). However, programmes for the prevention of mother-to-child transmission (PMTCT) have already reduced this rate by on average about a quarter in high prevalence countries and could feasibly bring it down to less than 8 per cent even in low-income countries (Mahy, Stover, Kiragu *et al*. 2010). The reduction in the fertility of infected women is typically about 25 per cent in populations in which few adults receive treatment with antiretroviral drugs but it is unclear how this may change in future (Chen and Walker 2010). Thus, for the reports of respondents born before PMTCT interventions became widespread, a suitable adjustment might be:[1]

$$ {}_5S_x = \frac{1-\frac{P}{3}}{1+\frac{P}{3}}\,{}_5S_x^*. $$

If the prevalence of HIV infection among women attending antenatal clinics at the time was 7 per cent, this becomes

$$ {}_5S_x = \frac{1-0.0233}{1+0.0233}\,{}_5S_x^* = 0.954\,{}_5S_x^*. $$

Correcting the reported proportions of respondents with living fathers is more difficult. The appropriate downward adjustment depends on the proportion of infected men with infected partners. Unfortunately, this statistic is often unknown and the evidence suggests that it varies substantially between populations, so using an average value may give rather imprecise results (de Walque 2007; Freeman and Glynn 2004; Kaiser, Bunnell, Hightower *et al*. 2011). Moreover, it has not been demonstrated that a crude correction factor that ignores age-related differences in the various parameters performs adequately for data on fathers. Nevertheless, the adjusted proportions will usually be less biased than unadjusted ones in populations with severe epidemics. If *w* denotes the proportion of men with infected partners, the appropriate adjustment factor is

$$ {}_5S_x = (1-(1-(1-h)F)(1-w)P^*)\,{}_5S_x^* $$

where $P^*$ denotes a population-based estimate of HIV prevalence among men, in contrast to the antenatal clinic-based estimate for women. For example, for $h = 1/3$, $F = 0.75$, $w = 0.5$ and $P^* = 0.05$, one obtains

$$ \begin{aligned} {}_5S_x &= (1-(1-(1-0.3333)0.75)(1-0.5)0.05)\,{}_5S_x^* \\ &= 0.9875\,{}_5S_x^*. \end{aligned} $$

For respondents aged less than 15 years, the assumption that all parents who were infected when the respondents were born have already died is unrealistic even in data sets collected before the implementation of mass treatment programmes. About half of this group of parents are alive for the 5–9 year old respondents and perhaps a quarter for 10–14 year old respondents. Using these figures, the correction (i.e. one minus the correction factor) made to the data on 5–9 year old respondents should be halved and that for those aged 10–14 reduced by a quarter.[2]

***Revised coefficients for estimating women's mortality***

Timæus and Nunn (1997) also develop a new set of regression coefficients for estimating life table survivorship from proportions of respondents with living mothers

1 This correction factor is somewhat larger than that suggested as a default by Timæus and Nunn (1997) as, on the basis of data available at that time, they adopted as typical values of the parameters $F = 0.8$ and $h = 0.25$.

2 Developing Timæus and Nunn's (1997) argument further leads to the result that one should reduce the size of the adjustment made to data on the first two age groups by multiplying the initial correction factors by:

$$ (1 + P(d-F)/F)/(1-(1-(1-h)d)P) $$

where *d* denotes the estimated ratio of the proportions of the infected to the uninfected parents that are still alive defining their HIV status at the time of the respondents' birth. The impact of this procedure differs little from that of simply reducing the complement of the correction factor by the proportion of parents of respondents in the age group estimated to remain alive.

in populations experiencing severe HIV epidemics (see Table 22.7). They are intended for use in populations in which the prevalence of HIV infection among adults is 5 per cent or greater. The regression model is the same as for the standard method and the coefficients only differ from the standard ones for respondents aged less than 30:

$$ {}_{n}p_{25} = a(n) + b(n)\bar{M}^{f} + c(n)\,{}_{5}S_{n-5}\,. $$

Equivalent coefficients for the estimation of men's mortality have not been developed because little is known about how age patterns of HIV incidence and fertility interrelate among men.

*TABLE 22.7 Coefficients for the estimation of women's survivorship from the proportions of respondents with living mothers in populations experiencing a generalized HIV epidemic*

| *n* | *a*(*n*) | *b*(*n*) | *c*(*n*) |
|---|---|---|---|
| 10 | –0.3611 | 0.00125 | 1.2974 |
| 15 | –0.4030 | 0.00222 | 1.3732 |
| 20 | –0.2120 | 0.00372 | 1.1342 |
| 25 | –0.2389 | 0.00586 | 1.1131 |
| 30 | –0.2513 | 0.00885 | 1.0223 |

Source: Timæus and Nunn (1997)

More information is available now than in the 1990s on age patterns of HIV incidence, survival post-infection, and the age pattern of fertility reduction among the HIV-infected. Thus, the coefficients in Table 22.7 could undoubtedly be refined. Nevertheless, they are unlikely to be severely biased. However, they may be inappropriate for use in populations in which treatment of HIV-infected adults with antiretroviral drugs has become widespread.

### *Estimation of the level and trend in mortality*

The biggest challenge involved in estimating adult mortality from orphanhood data in populations with generalized AIDS epidemics is that no adequate systems of model life tables exist with age patterns of mortality typical of a series of populations with AIDS epidemics of differing degrees of severity. Because AIDS deaths are concentrated among adults aged between about 25 and 55, changes in the level of mortality resulting from the spread in HIV also radically change the age pattern of mortality. This greatly reduces the confidence with which one can extrapolate from measures of survivorship for different age ranges to a common index of mortality. Moreover, as the incidence of HIV infection in most affected populations first grew very rapidly and is now decreasing, it may be appropriate to use different models for different cohorts of respondent. The workbook (see website) incorporates mortality standards devised for countries experiencing moderately severe HIV epidemics. However, these will not be appropriate either for countries where the prevalence of HIV infection in adults is just a few percentage points or countries with severe HIV epidemics. They may not even be appropriate for all countries with moderately severe HIV epidemics.

### *Illustrative application*

The problem that the analyst faces is illustrated by the estimates of adult survivorship made for Kenya from the 1999 Census data on orphanhood that are shown in Figure 22.2. The standard estimates, made without recourse to any of the adjustments described in this section and extrapolated to a common index of mortality using the United Nations General models, are shown using unbroken lines. They suggest that the probability of dying between ages 30 and 60 ($_{30}q_{30}$) in Kenya rose steadily from about 1987 to 1996 for both men and women, increasing by about 10 percentage points during this decade in each case.

The 'AIDS standard' estimates were produced in exactly the same way except that they were converted into estimates of $_{30}q_{30}$ using standards that incorporate mortality from AIDS. These results look very different: they suggest that the mortality of men continued to decline slowly in Kenya until the early 1990s but that otherwise adult mortality stagnated.

The 'AIDS' estimates show the additional effect of adjusting the proportions for AIDS-related selection bias and, for the most recent estimate for women, using the coefficients intended for populations in which the prevalence of HIV infection among adults exceeds 5 per cent. For 5–9 year old respondents, both adjusting for selection bias and using the new coefficients push up the estimate of women's mortality substantially. The HIV epidemic in Kenya in the 1990s was not severe enough, however, for the adjustments to make much difference to any of the other estimates.

To use the UN General model to estimate $_{30}q_{30}$ in Kenya for the 1990s amounts, in effect, to assuming that the rises in mortality from AIDS occurring among younger adults were matched by comparable rises in mortality among middle-aged individuals. This seems unlikely. The estimates of $_{30}q_{30}$ made using the AIDS standards, in contrast, imply that any

rise in the mortality of young adults was more than offset by continuing declines in the death rates of middle-aged adults in the late 1980s and more-or-less exactly offset by such declines till about 1993. However, few of the parents of respondents who were themselves in their thirties in 1999 are likely to have become infected with HIV and so using the AIDS standard to make the early estimates of ${}_{30}q_{30}$ also seems inappropriate. Perhaps the most likely scenario is that the probability of dying between ages 30 and 60 stagnated between the late-1980s and mid-1990s and then began to rise, though probably somewhat less abruptly than is indicated by the 'AIDS' series for women.

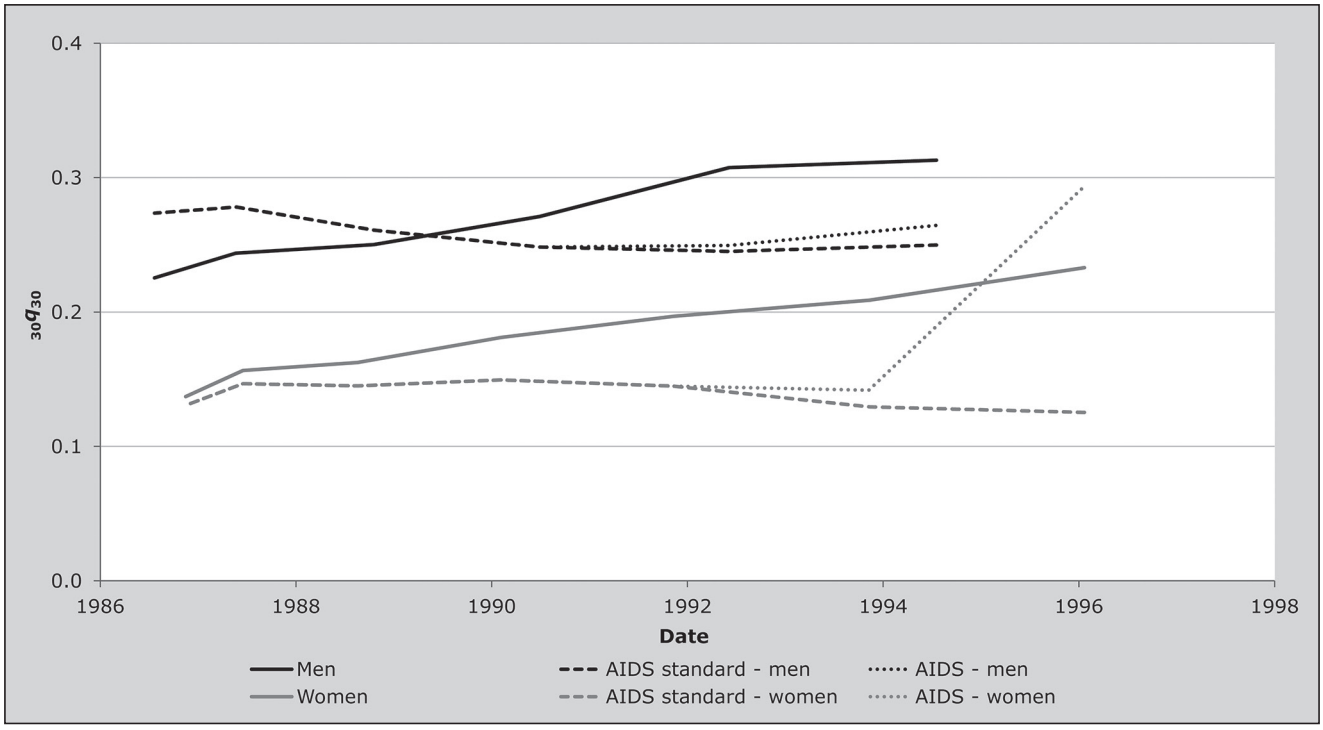
The crucial point to emphasize is that it is impossible to determine from these data exactly what happened to adult mortality in Kenya in the 1990s. The only estimates that are fairly reliable are those for the late 1980s. These results are based on age groups for which only limited adjustment of the proportions of parents that are alive is needed to estimate ${}_{30}q_{30}$. Thus, the series based on different standards intersect at this time.

**Orphanhood before and since marriage**

Methods exist for estimating adult mortality from orphanhood that can be used when supplementary questions are asked in a single survey about the timing of the deaths of parents relative to first marriage (Timæus 1991). Two aims underlie these methods. The first is to produce methods with a more specific and up-to-date time reference than those yielded by the original method. The second is to develop methods that are less subject to bias due to under-reporting of orphanhood at young ages.

Marriage is an event that distinguishes, for each age group of respondents, more recent parental deaths from those that occurred longer ago. While the information on the timing of deaths is less precise than that yielded by direct questions about the date of death of parents, it may be more accurately reported. Even if respondents cannot remember exactly when their parents died, it seems likely that nearly all of them will be able to report the timing of parental deaths relative to their first marriage, another event of major significance in their lives.

***FIGURE 22.2*** *Trends in the probability of dying between exact ages 30 and 60 estimated from orphanhood, Kenya, 1999 Census*



Some 15 inquiries conducted as part of Phase 1 of the Demographic and Health Surveys programme collected data on the relative timing of the deaths of parents and first marriage. Unfortunately, few or no surveys have done so more recently. Data on the timing of first births relative to the birth of respondents' first child could be analysed in exactly the same way and might be more representative of the mortality of all parents in populations in which many people never marry or in which getting married is thought of as a process rather than an event that occurs on a well-defined date.

Because men and women marry (and have their first births) at different ages, the reports of male and female respondent should be analysed separately. The estimation coefficients discussed here were developed primarily for the analysis of data supplied by female respondents (Timæus 1991) but could also be used for the analysis of data supplied by men if their mean age at marriage is less than 25. For respondents whose age exceeds the mean age at marriage, the estimates are robust to any characteristics of the distribution of ages of marriage other than its mean.

### *Orphanhood since first marriage*

The proportion of mothers that have remained alive since respondents married is closely related to the probability of surviving from the sum of the period mean age at childbearing and the cohort mean age at marriage until the sum of the mean age at childbearing and the current age of the respondents. Estimates made from data on orphanhood since marriage will measure more recent mortality than those based on respondents' lifetime experience of orphanhood. In addition, because parental deaths since marriage must have occurred when respondents were sufficiently old to remember them clearly, such data could be less subject to reporting errors than those concerning the overall level of orphanhood.

The earliest central age of respondents ($n$) for which one can estimate a survivorship ratio, ${}_{n}p_{b}$, from data on orphanhood since marriage is 30 years. For women, to preserve a close relationship between maternal survival and life table survivorship, the latter is measured from a base age, $b$, of 45 years. Thus, the model used to estimate life table measures from the survival of mothers since the respondents married is

$$ {}_{n-20}p_{45} = a(n) + b(n)\bar{M}^{f} + c(n)\bar{m} + d(n)\,{}_5S_{n-5}\big/{}_5S^{m}_{n-5} + e(n)\,{}_5S_{n}\big/{}_5S^{m}_{n} $$

where $\bar{m}$ is the cohort mean age at marriage and ${}_5S^{m}_{n}$ is the proportion of respondents in age group $n$ to $n+5$ who at the time of their first marriage had living mothers.

The only difference in the equation for the estimation of men's mortality from paternal orphanhood since marriage arises from the fact that men tend to be older than women at the birth of their children. Therefore, survivorship is estimated from a base age that is 10 years greater. The estimates are made using

$$ {}_{n-20}p_{55} = a(n) + b(n)\bar{M}^{m} + c(n)\bar{m} + d(n)\,{}_5S_{n-5}\big/{}_5S^{m}_{n-5} + e(n)\,{}_5S_{n}\big/{}_5S^{m}_{n} $$

where the age at marriage referred to is still that of the respondents, who will usually be female, not the parents, who are now male.

Until most respondents have married, the relationship between life table survivorship and the proportions of parents alive is sensitive to the shape of the age distribution of first marriages. Thus, coefficients exist for estimating survivorship over the age ranges ${}_{10}p_{45}$ to ${}_{30}p_{45}$ for adult women and ${}_{10}p_{55}$ to ${}_{20}p_{55}$ for adult men (see Table 22.8).

*TABLE 22.8 Coefficients for estimating adult women's and men's mortality from orphanhood since first marriage*

| $n$ | $a(n)$ | $b(n)$ | $c(n)$ | $d(n)$ | $e(n)$ |
|---|---|---|---|---|---|
| **Adult women** | | | | | |
| 30 | 0.5617 | 0.00836 | –0.00261 | –1.1231 | 1.4199 |
| 35 | 0.0476 | 0.01396 | –0.00536 | –0.3916 | 1.1354 |
| 40 | –0.3715 | 0.01966 | –0.00744 | 0.5394 | 0.5286 |
| 45 | –0.6562 | 0.02587 | –0.00716 | 1.0208 | 0.1789 |
| 50 | –0.8341 | 0.03045 | –0.00561 | 1.1898 | 0.0541 |
| **Adult men** | | | | | |
| 30 | 0.0676 | 0.01588 | –0.00633 | –1.2070 | 1.8284 |
| 35 | –0.5459 | 0.02273 | –0.01083 | –0.2509 | 1.3867 |
| 40 | –0.8674 | 0.02622 | –0.01135 | 0.6057 | 0.7198 |

Source: Timæus (1991)

### *Orphanhood before first marriage*

The proportion of women with living mothers at marriage approximately equals the life table probability of surviving from the mean age at childbearing of the mothers to that age plus the mean age at first marriage of their daughters. If the data are not biased by the adoption effect, this variant on the orphanhood method has two valuable characteristics. First,

it measures mortality over a limited and fairly clearly defined interval of time and range of ages. Secondly, the estimates are capable of extending the time series of mortality estimates provided by the orphanhood method backward to at least 30 or 35 years before the data were collected.

For orphanhood before marriage an interaction term between the mean age at marriage and proportion orphaned improves the fit of the model. Thus, the probability of surviving from age 25 to age 45 for women can be estimated from orphanhood before marriage using the regression equation:

$$ {}_{20}p_{25} = a(n) + b(n)\overline{M}^{f} + c(n)\overline{m} + d(n)\,{}_{5}S_{n}^{m} + e(n)\overline{m}\,{}_{5}S_{n}^{m}. $$

The coefficients of this equation for the different age groups defined by *n* are presented in Table 22.9.

The same considerations apply to the estimation of adult male mortality from paternal orphanhood before marriage. Estimates of the probability of surviving from age 35 to age 55 are made from

$$ {}_{20}p_{35} = a(n) + b(n)\overline{M}^{f} + c(n)\overline{m} + d(n)\,{}_{5}S_{n}^{m} + e(n)\overline{m}\,{}_{5}S_{n}^{m}. $$

These coefficients are also presented in Table 22.9.

*TABLE 22.9 Coefficients for estimating adult women's and men's mortality from orphanhood before first marriage*

| *n* | *a*(*n*) | *b*(*n*) | *c*(*n*) | *d*(*n*) | *e*(*n*) |
|---|---|---|---|---|---|
| **Adult women** | | | | | |
| 30 | –0.9607 | 0.00418 | 0.04466 | –0.04291 | 1.8178 |
| 35 | –0.9921 | 0.00429 | 0.04700 | –0.04501 | 1.8428 |
| 40 | –1.0129 | 0.00433 | 0.04822 | –0.04611 | 1.8607 |
| 45 | –1.0206 | 0.00434 | 0.04861 | –0.04648 | 1.8680 |
| **Adult men** | | | | | |
| 30 | –1.2719 | 0.01060 | 0.04480 | –0.04007 | 1.8383 |
| 35 | –1.2977 | 0.01068 | 0.04652 | –0.04124 | 1.8530 |
| 40 | –1.3203 | 0.01070 | 0.04769 | –0.04225 | 1.8726 |
| 45 | –1.3232 | 0.01070 | 0.04783 | –0.04238 | 1.8753 |

Source: Timæus (1991)

For orphanhood before marriage, the final sets of coefficients presented for maternal and paternal orphanhood before marriage in Table 22.9 refer to an age group in which first marriage is almost complete. These coefficients can be used to estimate mortality from the reports of any age group of respondents aged 40 and above. Only the accuracy of reporting about parental deaths and respondents' own ages impose an upper age limit on the data on orphanhood before marriage that can be used to estimate mortality.

***Time location of the estimates***

Like those from lifetime orphanhood, estimates made from the data supplied by different age cohorts of respondents, about orphanhood before and after marriage, reflect mortality over varying and ill-defined periods of time. For orphanhood before marriage, the time reference of the mortality measures is the product of the distribution of intervals between parental deaths and marriage and the distribution of intervals between marriage and interview. To the degree of precision required, the time reference of maternal orphanhood estimates is the average interval from orphanhood to first marriage plus the average interval from first marriage to interview. For cohorts of women who have nearly all married, the ages at which the parents are exposed to the risk of dying are concentrated between their mean age at childbearing and the sum of that age and their daughters' mean age at first marriage. The time reference of the mortality measures can be calculated as

$$ T = (N - \overline{m}) + {}_{\overline{m}}g_{\overline{M}}. $$

The second term on the right-hand side of Equation 5 can be estimated using the procedure explained with reference to the basic method (see Step 4). As for lifetime orphanhood, the time reference of estimates of male mortality should allow for deaths between conception and birth and are calculated using ${}_{\overline{m}+0.75}g_{\overline{M}-0.75}$.

For orphanhood since marriage, the age at which parents enter exposure to the risk of death is the product of their distribution of ages at childbearing and their daughters' ages at marriage. For age groups that have largely completed marriage, one can estimate the mean age at which exposure starts as the sum of the means of these two distributions. The parents' exposure continues till the time of interview, $N - \overline{m}$ years later. Thus, the time references of the mortality measures are

$$ T = {}_{N-\overline{m}}g_{\overline{M}+\overline{m}}. $$

Because the age range over which fathers are exposed to the risk of death commences well after the birth of their daughters, this equation is appropriate for estimates of both men's and women's mortality.

***Illustrative application***

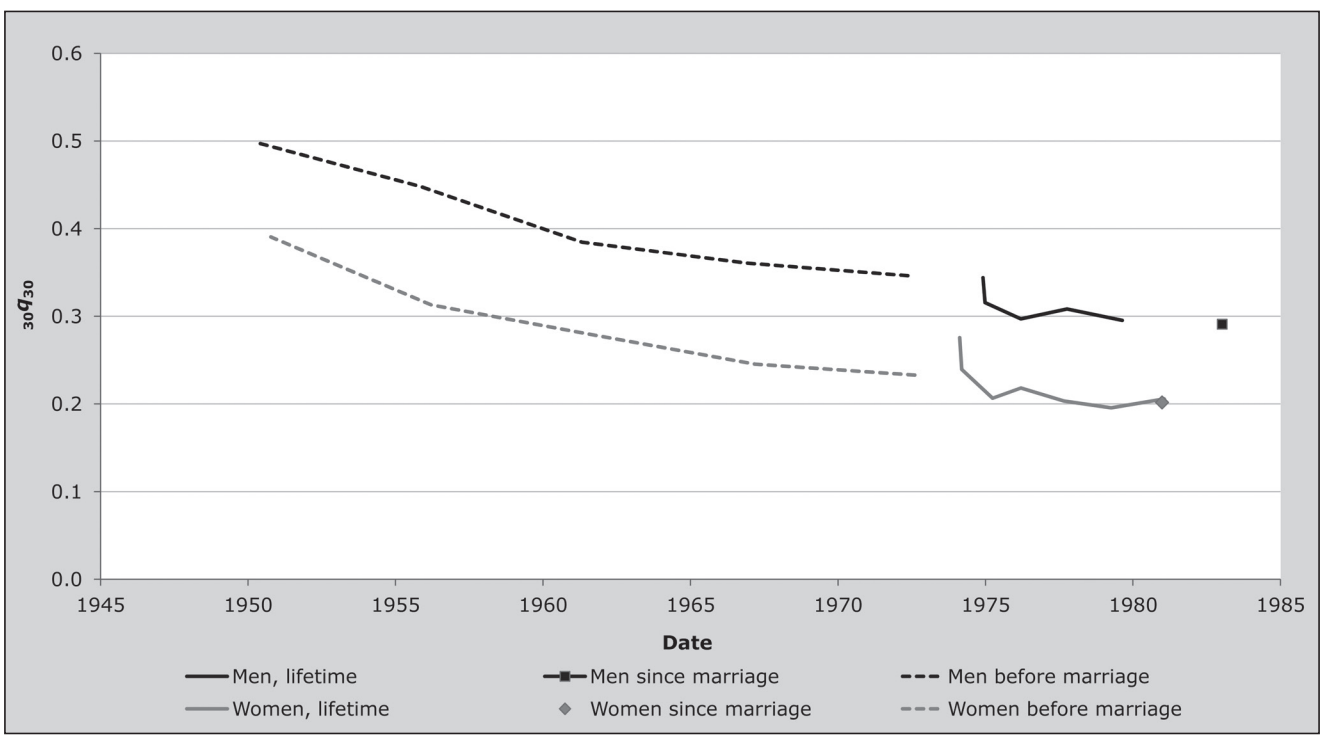
Figure 22.3 portrays the results of the application of the orphanhood before and since first marriage method for estimating adult mortality to data collected from women aged 15 to 49 in the 1988 Demographic and Health Survey (DHS) of Egypt. The first striking feature of these results is that the information on whether the older respondents were orphans when they first married collected in 1988 generates a series of estimates of adult mortality that extends back to the early 1950s. Second, the additional sets of estimates produced by partitioning the orphaned women into those who were orphaned before they first married and those who have been orphaned since, tie in consistently with those produced from data on lifetime orphanhood. These results offer no evidence that the underreporting of orphanhood due to the adoption effect is biasing down the most recent estimates based on lifetime orphanhood or those based on orphanhood before first marriage.

According to these results, a steady, gradually decelerating drop in adult mortality occurred in Egypt between the early-1950s and mid-1980s. The probability of dying between exact ages 30 and 60 fell by about 200 per thousand over this period for both men and women from a very high level in the early 1950s to about 200 per thousand for women and 300 per thousand for men in the early 1980s. Even these most recent estimates represent rather high mortality. The 100 per thousand difference between the probabilities of dying of men and women changed hardly at all over the period. The orphanhood estimates of adult mortality for men, in particular, are substantially higher than those based on other sources (UN Population Division 2011).

**Questions about the timing of parental deaths**

A further extension to the orphanhood method is to ask respondents whose parents have died about when the death occurred (Chackiel and Orellana 1985). If the dates on which parents died are reported with reasonable accuracy, this enables the analyst to distinguish recent from more distant parental deaths and obtain more up-to-date estimates of mortality. The best way to analyse such data is to use the information on dates of death to reconstruct the proportion of respondents who had living parents five and

*FIGURE 22.3 Trends in the probability of dying between exact ages 30 and 60 estimated from orphanhood, Egypt, 1988 DHS*



ten years earlier. From these successive cross-sections, one can construct synthetic cohort measures of parental survival that are formally identical to those generated from data collected in a series of separate inquiries. Therefore, methods for the analysis of these data are discussed jointly with the analysis of orphanhood data from multiple inquiries.

## FURTHER READING AND REFERENCES

The orphanhood method is discussed in all the classic manuals on indirect estimation (Sloggett, Brass, Eldridge *et al.* 1994; UN Population Division 1983) but, with the exception of the United Nations' manual on estimating adult mortality (UN Population Division 2002), these manuals give emphasis to the older variant of the method that uses weighting factors to produce life table indices, rather than the regression-based method normally used today. Although regression-based methods for women had been proposed previously (Hill and Trussell 1977; Palloni and Heligman 1985), regression methods for estimating men's mortality were first developed by Timæus (1992). His article also surveys earlier contributions to the literature and discusses the theoretical basis of the method.

Methods for estimating adult mortality from orphanhood before and after marriage are described in Timæus (1991). Procedures for estimating women's mortality in populations experiencing generalized HIV epidemics were developed by Timæus and Nunn (1997).

Luy (2012) has proposed a modified orphanhood method intended for the study of socio-economic differentials in adult mortality in countries that already have accurate aggregate information on adult mortality. His method determines the relationship between the proportion of survey respondents reporting living parents and life table survivorship from empirical statistics on the population concerned instead of using demographic models. Thus, Luy's method tailors the estimation process to the context. Few low- and middle-income countries have sufficiently accurate national data for this to be feasible, though aspects of the approach may be of relevance in some circumstances.

Methods that exploit the additional analytic opportunities that arise when questions about orphanhood have been asked in two or more successive inquiries in the same population are discussed in Chapter 26.


Blacker JGC. 1984. "Experiences in the use of special mortality questions in multi-purpose surveys: the single-round approach," in *Data Bases for Mortality Measurement*. New York: United Nations, pp. 79–89. http://www.un.org/esa/population/publications/UN_1984_Data_Bases_for_Mortality_Measurement/UN1984_Data_Bases.htm

Blacker JGC and JM Gapere. 1988. "The indirect measurement of adult mortality in Africa: results and prospects," in *African Population Conference, Dakar, 1988*. Liège: International Union for the Scientific Study of Population,Vol. 2:3.2.23–38.

Brass W. 1971. "On the scale of mortality," in Brass, W (ed). *Biological Aspects of Demography*. London: Taylor and Francis, pp. 69–110.

Brass W. 1975. *Methods for Estimating Fertility and Mortality from Limited and Defective Data.* Chapel Hill: International Program of Laboratories for Population Statistics.

Brass W. 1985. *Advances in Methods for Estimating Fertility and Mortality from Limited and Defective Data.* London: London School of Hygiene & Tropical Medicine.

Brass W and EA Bamgboye. 1981. *The Time Location of Reports of Survivorship: Estimates for Maternal and Paternal Orphanhood and the Ever-widowed*. London: London School of Hygiene & Tropical Medicine.

Brass W and K Hill. 1973. "Estimating adult mortality from orphanhood," in *International Population Conference, Liège, 1973*. Liège: International Union for the Scientific Study of Population, Vol. 3:111–123.

Chackiel J and H Orellana. 1985. "Adult female mortality trends from retrospective questions about maternal orphanhood included in censuses and surveys," in *International Population Conference, Florence, 1985*. Liège: International Union for the Scientific Study of Population, Vol. 4:39–51.

Chen W-J and N Walker. 2010. "Fertility of HIV-infected women: insights from Demographic and Health Surveys", *Sexually Transmitted Infections* **86**(Suppl 2):ii22–ii27. doi: 10.1136/sti.2010.043620

De Cock KM, MG Fowler, E Mercier, I de Vincenzi *et al.* 2000. "Prevention of Mother-to-Child HIV Transmission in Resource-Poor Countries", *Journal of the American Medical Association* **283**(9):1175–1182. doi: 10.1001/jama.283.9.1175

de Walque D. 2007. "Sero-discordant couples in five African countries: implications for prevention strategies", *Population and Development Review* **33**(3):501–523. doi: 10.1111/j.1728-4457.2007.00182.x

Freeman EE and JR Glynn. 2004. "Factors affecting HIV concordancy in married couples in four African cities", *AIDS* **18**(12):1715–1721. doi: 10.1097/01.aids.0000139075.13906.2f

Henry L. 1960. "Mesure indirecte de la mortalité des adultes", *Population* **15**:457–466.

Hill K. 1984. "An evaluation of indirect methods for estimating mortality," in Vallin J, JH Pollard and L Heligman (eds). *Methodologies for the Collection and Analysis of Mortality Data*. Liège, Belgium: Ordina Editions, pp. 145–176.

Hill K and TJ Trussell. 1977. "Further developments in indirect mortality estimation", *Population Studies* **31**(2):313–334. doi: 10.2307/2173920

Kaiser R, R Bunnell, A Hightower, AA Kim *et al.* 2011. "Factors Associated with HIV Infection in Married or Cohabitating Couples in Kenya: Results from a Nationally Representative Study", *PLoS ONE* **6**(3):e17842. doi: 10.1371/journal.pone.0017842

Luy M. 2012. "Estimating mortality differences in developed countries from survey information on maternal and paternal orphanhood", *Demography* **49**(2):607–627. doi: 10.1007/s13524-012-0101-4

Mahy M, J Stover, K Kiragu, C Hayashi *et al.* 2010. "What will it take to achieve virtual elimination of mother-to-child transmission of HIV? An assessment of current progress and future needs", *Sexually Transmitted Infections* **86**(Suppl 2):ii48–ii55. doi: 10.1136/sti.2010.045989

Palloni A and L Heligman. 1985. "Re-estimation of structural parameters to obtain estimates of mortality in developing countries", *Population Bulletin of The United Nations* **18**:10–33.

Palloni A, M Massagli and J Marcotte. 1984. "Estimating adult mortality with maternal orphanhood data: analysis of sensitivity of the techniques", *Population Studies* **38**(2):255–279. doi: 10.1080/00324728.1984.10410289

Sloggett A, W Brass, SM Eldridge, IM Timæus, P Ward and B Zaba. 1994. *Estimation of Demographic Parameters from Census Data.* Tokyo, Japan: United Nations Statistical Institute for Asia and the Pacific.

Timæus I. 1986. "An assessment of methods for estimating adult mortality from two sets of data on maternal orphanhood", *Demography* **23**(3):435–450. doi: 10.2307/2061440

Timæus IM. 1991. "Estimation of adult mortality from orphanhood before and since marriage", *Population Studies* **45**(3):455–472. doi: 10.1080/0032472031000145636

Timæus IM. 1992. "Estimation of adult mortality from paternal orphanhood: a reassessment and a new approach", *Population Bulletin of the United Nations* **33**:47–63.

Timæus IM and AJ Nunn. 1997. "Measurement of adult mortality in populations affected by AIDS: an assessment of the orphanhood method", *Health Transition Review* **7(Suppl 2)**:23–43.

UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

UN Population Division. 2002. *Methods for Estimating Adult Mortality.* New York: United Nations, Department of Economic and Social Affairs, ESA/P/WP.175. http://www.un.org/esa/population/techcoop/DemEst/methods_adultmort/methods_adultmort.html

UN Population Division. 2011. *World Mortality Report 2007.* New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/289. http://www.un.org/esa/population/publications/worldmortalityreport2007/WORLD%20MORTALITY%20REPORT.PDF

# Chapter 23 Indirect estimation of adult mortality from data on siblings

*Ian M Timæus*

## DESCRIPTION OF METHOD

This method estimates adult mortality indirectly from data supplied by adults on the survival of their adult siblings (that is brothers and sisters). These data are tabulated by the age group of the respondents. Mortality can be estimated from them without requiring respondents to recall the dates when deaths occurred or the ages at death of deceased individuals. Information on the survival of brothers is used to estimate the mortality of men and information on the survival of sisters to estimate the mortality of women.

Respondents often fail to report – and may not know about – siblings who died before or during the few years after their own birth. The impact of this bias can be reduced greatly, however, by only including siblings who survived to age 15 in the analysis. In order to apply the method, a census or survey must have asked adult respondents (for example, those aged 15 to 49) how many of their sisters and/or brothers survived to the age of 15 and how many of them are still alive. Many surveys only collect information on siblings from women but data supplied by male respondents can be analysed using exactly the same methods.

Respondents' siblings are approximately the same age, on average, as the respondents. Thus, the proportion of the siblings who survived to age 15 who are still alive is a good estimator of the conditional probability life table of surviving from age 15 to the current age of the respondents.

If mortality has changed over time, the estimated survivorship ratios reflect the mortality rates that have affected each cohort at a range of ages and dates. A 'time location' method has been developed that estimates how many years prior to the inquiry each cohort ratio equalled period survivorship. These intervals increase with the age of respondents, ranging from about 3 to 13 years before the collection of the data. Thus, if the survivorship ratios estimated from the reports of different age groups of respondent are translated into a common index of mortality in adulthood using a 1-parameter system of model life tables, these statistics will refer to different dates and can be used to infer the broad trend in mortality over time.

One advantage that sibling methods have over questions about household deaths is that only censuses or unusually large surveys can capture information on enough deaths in households in the year before the inquiry to yield mortality estimates that are sufficiently precise to be useful. Because respondents report on several siblings, on average, and the estimates are based on all exposure to risk at age 15 or more, estimates can be made from data on siblings in smaller inquiries. Nevertheless, all methods for the estimation of adult mortality require data on several thousand households. Moreover, the estimation procedure does not assume a population closed to migration. However, the results from the method will not be representative for small states or sub-national areas in which a substantial proportion of the population are in-migrants or have emigrated.

## DATA REQUIREMENTS AND ASSUMPTIONS

### Tabulations of data required

To estimate the mortality of adult women, respondents aged 15 to 49 should be asked how many of their sisters lived to age 15 and how many of these sisters are still alive. From tabulations of the answers to these two questions by the age group of the respondent, one can calculate:

- The proportion still alive of sisters who were alive on their 15th birthday by five-year age group of respondent. (Those who did not answer either question should be excluded from the calculations.)

To estimate the mortality of adult men, respondents aged 15 to 49 should be asked how many of their brothers lived

to age 15 and how many of these brothers are still alive. From tabulations of the answers to these two questions by the age group of the respondent, one can calculate:

- The proportion still alive of brothers who were alive on their 15th birthday by five-year age group of respondent. (Those who did not answer either question should be excluded from the calculations.)

The tabulations of numbers of siblings reaching age 15 and still alive should *exclude* the respondent himself or herself. (Of course, this is always the case when respondents report on siblings of the opposite sex.) This requirement is explained in the discussion of the important assumptions made by the method.

Tables on respondents' own-sex siblings (i.e. women's sisters and men's brothers) should be weighted only by any sample or design weights provided with the data. Tables on respondents' opposite sex siblings (i.e. women's brothers and men's sisters) should be further weighted by the inverse of the number of surviving own-sex siblings of the individual respondent making the reports. This requirement is also explained in the discussion of the important assumptions made by the method.

To eliminate ambiguities related to polygynous marriage and to remarriage, interviewers in most inquiries are instructed that 'siblings' means children born to the same mother. Whether or not this has been done, the reports can usually be accepted as they are. So long as respondents have the same group of relatives in mind when answering the question about siblings that are still alive as they did when answering the preceding question about siblings who survived to age 15, for the purpose of estimating mortality it is immaterial exactly who their parents are.

If both men and women have been asked the relevant questions, their responses should usually be tabulated separately so that the two sets of data can be checked against each other.

Many Demographic and Health Surveys collect full sibling histories from all women aged 15 to 49. These histories ask each respondent for the name, sex, age, survival status and, if dead, age at and year of death for each of their siblings born to the same mother. All-cause death rates for men and women should usually be calculated from such histories by the direct sibling method. The reporting of recent deaths of siblings is believed to be more complete than that of more distant deaths and the direct method allows one to restrict the analysis to data on the years immediately prior to the conduct of the survey. It is straightforward, however, to determine the summary counts required by the indirect method from the full sibling histories. If reporting of siblings is fairly complete, but the reporting of the ages and dates of death of siblings is very poor, the indirect approach might yield more reliable results than the direct one.

**Important assumptions**

An inherent limitation of sibling-based methods for measuring adult mortality is that they underestimate mortality insofar as mortality clusters within sibships (i.e. sets of brothers and/or sisters born to the same mother). Clustering occurs whenever deaths are more concentrated in a small proportion of sibships than would be expected by chance and results from inter-sibship heterogeneity in individuals' risk of dying (Zaba and David 1996). It causes downward bias in the mortality estimates simply because fewer members of a high mortality sibship than a low mortality sibship of the same size remain alive to answer questions about their siblings. It is impossible to correct fully for this because, at the extreme, sets of siblings whose members have all died are not reported on at all. There is no way of knowing how many of these sibships existed or what their sizes were.

Estimates of mortality trends will be biased as a result if the extent to which clustering of mortality within sibships varies with age. For example, if characteristics shared by sibs (e.g. genetic factors, early-life experiences, socio-economic status, life styles, and location) strongly influence the mortality of middle-aged adults, whereas mortality before age 40 has a large random component, estimates made for older respondents will underestimate mortality by more than those made from data supplied by younger respondents, producing a spurious impression of mortality increase over time.

The issue of bias related to multiple reporting of siblings has received substantial attention in the literature. The problem exists in survey as well as census data because the more times an individual would be reported in a census, the more likely they are to have a sibling who reports on them included in a probability sample.[1] Moreover, even in surveys,

1 Note that the issue of calculating the sampling error of the mortality estimates correctly in surveys in which some individuals are reported on more than once and mortality is clustered at the level of the sibship, is distinct from the issue being discussed here, which is that of bias in the central estimate of mortality.

potential exists for multiple responses about the same individual. For example, if two daughters of the same mother are interviewed in the same household, there will be multiple reports about other members of the sibship. The standard approach to analysis used, for example, in DHS reports is based on the events and exposure time of reported siblings, leaving out the exposure time of the (surviving) respondent herself. Events and exposure time are weighted only by the respondent's sample weight, not taking into account numbers of surviving potential respondents in the sibship.

Trussell and Rodriguez (1990) demonstrate mathematically that for groups of same-sex sibships with an identical underlying risk of dying, the standard calculation that excludes the respondent from both the denominator of the measures yields unbiased estimates of mortality. In effect, the reduction in the number of reports on dead people in the numerator that occurs because dead people cannot report on one another and the exclusion of the living respondents from the denominator offset each other precisely to give the correct risk for the sibships as a group.

The issue of the biases that could result from differential mortality by sibship size is bound up with the issue of multiple-reporting bias. It has attracted a lot of research interest because, unlike other factors that affect risk within sibships classified by sex and age of the respondent, each respondent's sibship size is known. If mortality does not vary with sibship size, the standard estimates are the same for both every size of sibship, including one-person sibships that are excluded from the analysis because the respondent has nobody to report on, and the population as a whole. Even if mortality varies by sibship size the standard estimates remain unbiased for each sibship size, as pointed out by Masquelier (2013). To obtain mortality estimates for the population though, one must reweight the estimates for sibships of different sizes by the prevalence of sibships of that size in the population. When respondents are reporting on their own sex, one can achieve this by dividing the proportion of respondents from surviving sibships of each size by the estimated proportion of siblings surviving from age 15 to that age group in all sibships of the same size. To do this for single-person sibships, their mortality has to be estimated by extrapolation from mortality in larger sibships.

Gakidou and King (2006) argue that, instead of the standard approach, the proportions dying should be estimated for sibships that include the surviving respondent but should always be weighted in addition by the likelihood that they will be reported – that is, by the inverse of the number of potential respondents surviving in the sibship. As in Masquelier's approach, an additional adjustment also must be made for sibships that go unreported because no member remains alive. In a multi-survey analysis of DHS full sibling histories, Obermeyer, Rajaratnam, Park *et al.* (2010) estimate that the effect of not adjusting for the likelihood of reporting may bias overall mortality estimates downward by as much as 20 per cent. Masquelier (2013), however, argues that Obermeyer and her co-authors reweighted the data inappropriately and, as a result, exaggerated the size of any bias. He emphasizes that, if one is going to reweight, it is important only to adjust for multiple reporting by siblings who survived to the initial age from which mortality is being measured. In addition though, he questions whether the observed variation in mortality by sibship size is always real. Instead, it may well be an artefact of greater omission of dead siblings in the histories reported for large sibships. Masquelier therefore recommends using the standard approach, without attempting to reweight the data to allow for differential mortality by sibship size. This is the approach that is adopted here.

When analysing reports made on the opposite sex (for example, responses made by women about their brothers), the issues are rather different. In this case, the respondent is not a member of the group that is exposed to the risk of dying. However, the standard calculation will still give biased results for the population as a whole if the mortality of siblings of one sex is associated with the number of siblings of the opposite sex that report on them. Thus, for reports on the opposite sex a clear case exists for weighting each report by the inverse of the respondent's number of surviving siblings of their own sex as suggested by Gakidou and King (2006). Of course, questions about siblings of the opposite sex cannot generate any information on those sibships whose members have no living siblings of the respondent's sex. Thus, adopting this approach is equivalent to assuming that the mortality of individuals in such sibships is the same as the mortality of the rest of the population. However, in surveys that collect data from both sexes, each sex supplies this information for the other and one can further weight the deaths and exposure reported by respondents by the inverse of the probability that siblings in each age group get reported on at all.

The adult sibling method estimates the trend in mortality from data supplied by different age groups of respondent: the older the respondent, the longer ago their brothers and sisters

died on average. In order to convert the series of measures of survivorship obtained indirectly from data on different age groups into a single indicator that can be compared over time, it must be assumed that the age pattern of mortality in adulthood is represented by the chosen standard life table. To estimate the time location of these measures, it is further assumed that mortality declined linearly in terms of that standard over the period being considered.

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Before starting the analysis, one should check how many respondents stated that they did not know how many of their siblings reached age 15 or how many of them are still alive, or who failed to answer the questions at all. Although the response rate on these questions is usually high, a few surveys have collected sufficiently incomplete data to suggest that non-response bias could be a substantial problem.

If both women and men have been asked the relevant questions, one useful check on the quality of the data on siblings is to assess how many siblings of each sex are reported, on average, by respondents of the other sex and whether the reported sex ratio at birth changes markedly as the respondents' age increases.

One should also compare the proportions dead reported by male and female respondents of the same age. The mortality of individuals of a particular sex as reported by their brothers should equal the mortality of the same individuals as reported by their sisters. If it does not, this may indicate significant bias in the proportions dead for one or both sexes as a result of the multiple reporting of some sibships and the fact that none of the sibs have survived to report on others. Alternatively, if the proportions reported by male and female respondents diverge as their age increases, this could reflect gender differences in patterns of age misreporting or the fact that the sex that reports fewer dead siblings (usually the men) is more likely to lose touch with their families of origin and tends to assume wrongly that some of their dead siblings remain alive.

## CAVEATS AND WARNINGS

- The original indirect sibling method was developed to estimate the survivorship of siblings from birth (Hill and Trussell 1977). Unfortunately, such reports are often very incomplete, particularly for siblings who died before or soon after the birth of the respondent. The indirect adult sibling method is recommended instead, but can only produce probabilities of surviving from age 15 to ages later in adulthood conditional on being alive at exact age 15. To produce a complete life table, one has to estimate survivorship from birth to age 15 using estimates based on another source of data.
- Deaths of siblings do not occur at one point in time but may have occurred at any time between the respondents' 15th birthday and when they were interviewed. Therefore, applications of the indirect sibling method can only indicate the smoothed trend in adult mortality and will fail to capture short-term mortality crises or abrupt reversals in the trend in mortality such as those resulting from AIDS after the onset of a generalized HIV epidemic.
- The most up-to-date mortality estimates that can be produced using data on adult siblings are made from information supplied by respondents aged 20–24. Typically, however, only a few per cent of their brothers and sisters have died and the resulting estimates can have wide confidence intervals even in large sample surveys.
- The indirect procedure for estimating adult mortality from information on adult sibling does not involve the assumption that the population is closed to migration. Nevertheless, it can be difficult to interpret sibling-based estimates of adult mortality for sub-national geographic units, such as urban and rural areas or districts, or for respondents with particular socio-economic characteristics. This is because, although siblings usually share the same ethnic identity, many of the respondents' siblings will live in different places from the respondents themselves and their socio-economic characteristics may differ from those of the respondents. Estimates for sub-national populations are also likely to have very large sampling errors.

## APPLICATION OF METHOD

### *Step 1: Calculate the conditional life table survivorship ratios*

The procedure for estimating life table indices from the proportions of siblings that are alive is identical no matter whether one is analysing data on brothers, sisters or siblings of both sexes and irrespective of whether the respondents are men, women or of both sexes. The accompanying Excel workbook (see website) contains panels for estimating both men's and women's survivorship from data on male respondents, female respondents and respondents of both sexes. Either the number of brothers or sisters alive at age

15 and the number of brothers or sisters still alive by five-year age group of respondent or the proportions of those brothers/sisters who were alive at 15 that are still alive by five-year age group of respondent should be entered into the appropriate panels. Estimates are produced from the data on respondents aged 20 to 49 years.

Survivorship is estimated between exact age 15 years and $15+n$ years, where $n$ is the upper limit of each successive age group of respondents. The following regression equation and the coefficients shown in Table 23.1 are used:

$$ {}_{n-15}p_{15} = a(n) + b(n)\,{}_5S_{n-5} $$

where ${}_5S_{n-5}$ represents the proportion still alive of the siblings who were alive on their 15th birthday of respondents aged $n-5$ to $n$. For example, when $n$ is 25, life table survivorship is estimated over the 10-year age interval from exact age 15 to exact age 25 using data on survival of siblings supplied by respondents in the age group 20–24 years.

*TABLE 23.1 Coefficients for the estimation of adult survivorship from the proportions of respondents with living sisters or brothers*

| *n* | *a*(*n*) | *b*(*n*) |
|---|---|---|
| 25 | –0.0003 | 1.0011 |
| 30 | –0.1546 | 1.1560 |
| 35 | –0.1645 | 1.1660 |
| 40 | –0.1388 | 1.1406 |
| 45 | –0.1140 | 1.1168 |
| 50 | –0.1018 | 1.1066 |

Source: Timæus, Zaba and Ali (2001)

### *Step 2: Convert the survivorship ratios into estimates of the level of mortality*

To infer mortality trends from a series of survival ratios, ${}_{n-15}p_{15}$, obtained from different age groups of respondents and referring to different dates, it is necessary to convert them all into a common index of mortality that can be compared over time. This is done by fitting a 1-parameter relational logit model life table to each measure and obtaining the common mortality index from the model.

A wide range of indices have been used for this purpose, including the level parameters of various systems of model life tables, survivorship ratios such as ${}_{35}p_{15}$, which corresponds to the widest age range for which the adult siblings method yields measures, and life expectancy at age 15. Using the parameters of the models has the advantage of emphasizing that the full life table is being estimated by fitting a model, rather than measured directly. The measure of life expectancy summarizes survivorship across adulthood as a whole, while using survivorship avoids extrapolation into old age from measures for younger adults. Increasingly, in recent years, the estimates have been presented in terms of the probability of a 15-year old person dying before age 60, ${}_{45}q_{15}$, as this measure has found favour with several international agencies as a summary indicator of the mortality of young and middle-aged adults.

In the applications of the adult sibling method presented here the survivorship ratios are converted into values of $\alpha$, the level parameter of a system of relational logit model life tables, and then into estimates in these fitted models of the conditional probability of a 15-year old person dying by exact age 60 (${}_{45}q_{15}$), exact age 50 (${}_{35}q_{15}$) or exact age 40 (${}_{25}q_{15}$). The last of these indices lies in the middle of the series of ${}_{n-15}q_{15}$ values estimated using the regression models. The model life table parameter is calculated as

$$ \alpha = -\frac{1}{2}\ln\left(\frac{1+\dfrac{{}_{n-15}p_{15}}{l^s(n)} - \dfrac{1}{l^s(15)}}{1-{}_{n-15}p_{15}}\right) $$

where the estimates of ${}_{n-15}p_{15}$ come from Step 1 and the $l^s(x)$ values come from a standard life table. Thus, one obtains a series of values of $\alpha$ corresponding to the survivorship estimates made from data on the different age groups of respondents. Higher values of $\alpha$ correspond to higher mortality. Then for each $\alpha$,

$$ {}_xq_{15} = 1 - \frac{1+e^{2\left(\alpha+Y^s(15)\right)}}{1+e^{2\left(\alpha+Y^s(15+x)\right)}} $$

for $x$ = 25, 35, and 45.

The workbook (see website) can calculate $\alpha$ and the three summary probabilities of dying using either a standard from the General family of UN Population Division (1982) model life tables or one from any of the four families of Princeton model life tables (Coale, Demeny and Vaughan 1983). The standard life table should be chosen to have an age pattern of mortality within adulthood that resembles that of the population being studied. Another life table can be used as a standard if there is reason to believe that it resembles more closely the pattern of adult mortality in the population being studied. The most suitable standard may not be the family of models that best captures the relationship between child and adult mortality. If nothing is known about the age

pattern of mortality in adulthood, use of the United Nations General or Princeton West models is recommended.

***Step 3: Calculate the time location of the estimates***

Each survivorship ratio produced by the adult sibling survival method represents an average of the mortality prevailing during the period since the siblings' 15th birthdays. The duration of exposure of respondents' older siblings is longer than that of their younger siblings but averages out at about $n$–18.3 years (this is explained in a later section). The point of time at which the ratio equalled period survivorship depends on the level of mortality and can be estimated from the proportion of the siblings that are alive and the age of the respondents.

The only additional information required to calculate the dates to which each of the adult morality estimates refer is the date on which the inquiry that asked about the survival of brothers and sisters was conducted. This can be calculated as the average of the dates on which the interviews took place or taken as the mid-point of the period of fieldwork if exact dates of interview are not available.

The estimates are calculated as the date of fieldwork minus $T$, where $T$ can be calculated for each estimate using the following equation and the coefficients in Table 23.2:

$$T = a(n) - b(n)\ln\left({}_5S_{n-5}\right).$$

*TABLE 23.2 Coefficients for calculating the time location of estimates of adult survivorship from the proportions of respondents with living sisters or brothers*

| *n* | *a*(*n*) | *b*(*n*) |
|---|---|---|
| 25 | 3.23 | 1.12 |
| 30 | 5.46 | 1.95 |
| 35 | 7.52 | 2.78 |
| 40 | 9.38 | 3.62 |
| 45 | 11.00 | 4.45 |
| 50 | 12.32 | 5.28 |

Source: Timæus, Zaba and Ali (2001)

## WORKED EXAMPLE

This example uses data on the survival of brothers and sisters collected in the 2003 World Health Survey in Bangladesh (accessed 17 November 2012). This survey collected data from adult men and adult women on both their brothers and sisters. The reports made by women about their sisters are used to illustrate the calculations involved in the method.

***Step 1: Calculate the conditional life table survivorship ratios***

The number of sisters that respondents reported as having survived to their 15th birthdays and the number of these sisters that they reported to be still alive are shown in the second and third columns of Table 23.3. These numbers were tabulated using the standard approach, weighting the reports only by the survey weights. The proportions alive in the fourth column are calculated by dividing the counts in the third column by those in the second column. The survivorship ratios are presented in the sixth column of Table 23.3. They were estimated from the proportions in the fourth column using the regression coefficients shown in Table 23.1. For example, for respondents aged 25–29,

$${}_{10}p_{15} = -0.0003 + 1.0011 \times 0.9533 = 0.9541.$$

***Step 2: Convert the survivorship ratios into estimates of the level of mortality***

The seventh column of Table 23.3 contains values for $\alpha$, the level parameter of the system of relational logit model life tables based on a South Princeton model life table for both sexes with $e_0$=60. The eighth column of the table contains estimates of ${}_{25}q_{15}$ (the probability of a 15-year old person dying before their 40th birthday). For example, $\alpha$ is calculated from the estimate of ${}_{10}p_{15}$ for women as

$$\alpha = -\frac{1}{2}\ln\left(1 + \frac{\dfrac{0.9541}{0.8376} - \dfrac{1}{0.8557}}{1 - 0.9541}\right) = 0.517.$$

Having calculated $\alpha$, then the corresponding measure of ${}_{25}q_{15}$ is

$${}_{25}q_{15} = 1 - \frac{1 + e^{2(0.517 - 0.8899)}}{1 + e^{2(0.517 - 0.6902)}} = 0.136.$$

***Step 3: Calculate the time location of the estimates***

The dates in the ninth column of Table 23.3 were calculated by subtracting the time location of the estimates from the date of the 2003 World Health Survey in Bangladesh, which was taken as 30/04/2003. The time location of each estimate is calculated from the age group of the respondents and the proportions surviving using the equations in Table 23.2. For example, for the first estimate of sisters' survivorship, the calculation is:

$$T = 3.23 - 1.12 \times \ln\left(0.9533\right) = 3.28.$$

## DIAGNOSTICS, ANALYSIS AND INTERPRETATION

### Checks and validation

The number of respondents who stated that they did not know either how many of their siblings reached age 15 or how many of them are still alive or who did not answer the question at all should be checked before they are dropped from the analysis. If a lot of respondents failed to respond to these questions, the data supplied by those respondents who did answer them may not be representative of the population as a whole. Moreover, a high level of non-response may indicate that either the field staff or the respondents were experiencing difficulty with the questions. Thus, even when answers were supplied they may be rather unreliable. If a high level of item non-response exists, it can be illuminating to determine whether it is concentrated among a minority of field staff or a certain type of respondent.

If information about the survival of siblings has been collected from both men and women in a census or a large survey, tabulating the proportions of brothers and sisters alive separately for male and female *respondents* can be recommended in order to compare the consistency of their reports. While consistency of reporting does not guarantee accuracy, statistically significant differences between the proportions obtained from male or female respondents imply either that bias has been introduced into the estimates by the design of the question or that respondents of one or both sexes are answering the questions inaccurately. It is fairly common to find that male respondents report fewer siblings than female respondents and, in particular, fewer dead siblings. In other surveys, the two sexes may report similar numbers of siblings surviving to age 15 but that different numbers of them remain alive. The first type of discrepancy might result from differential age misreporting, but the second cannot.

Any bias due to clustering of mortality within families results in underestimates. Moreover, it seems unlikely that respondents invent siblings or report that their living siblings have died. Thus, the analysis should probably focus on the data for the sex that reports lower proportions of living siblings.

### Interpretation

The results of the example application of the indirect adult sibling method to data from the 2003 World Health Survey in Bangladesh are portrayed graphically in Figure 23.1. The life table indicator presented in this instance is the probability of dying between ages 15 and 40.

The proportions of women's sisters alive among those who lived to age 15 and of men's brothers surviving among those who lived to age 15 were tabulated using the standard approach, weighting the reports only by the survey weights. In contrast, the equivalent proportions of men's sisters and women's brothers alive were tabulated after further weighting each respondent's responses by the inverse of their number of surviving same-sex siblings.

It can immediately be seen that all four series of estimates fluctuate somewhat erratically and that all four series also tend to suggest that adult mortality rose in Bangladesh during the 1990s. This seems unlikely and may suggest that the estimates made from data on older respondents are biased downward by omission of dead siblings from the reports or by other biases.

The estimates of the mortality of women (i.e. the sisters'

*TABLE 23.3 Estimation of women's survivorship, the dates at which these estimates are located in time, and corresponding estimates of $\alpha$ and ${}_{25}q_{15}$, from the survival of adult sisters, Bangladesh, 2003*

| Age group | Sisters alive at 15 | Sisters still alive | Proportion alive (${}_5S_{n-5}$) | $n$ | $\frac{l(n)}{l(15)}$ | Level ($\alpha$) | Probability of dying (${}_{25}q_{15}$) | Date |
|---|---|---|---|---|---|---|---|---|
| 15–19 | 871.6 | 851.8 | 0.9773 | | | | | |
| 20–24 | 858.8 | 818.7 | 0.9533 | 25 | 0.9541 | 0.517 | 0.136 | 2000.0 |
| 25–29 | 964.7 | 901.1 | 0.9340 | 30 | 0.9251 | 0.535 | 0.139 | 1997.7 |
| 30–34 | 766.9 | 702.2 | 0.9156 | 35 | 0.9031 | 0.467 | 0.128 | 1995.6 |
| 35–39 | 626.4 | 554.4 | 0.8850 | 40 | 0.8706 | 0.473 | 0.129 | 1993.5 |
| 40–44 | 552.8 | 490.7 | 0.8877 | 45 | 0.8774 | 0.226 | 0.093 | 1991.8 |
| 45–49 | 495.9 | 401.5 | 0.8095 | 50 | 0.7940 | 0.436 | 0.124 | 1989.9 |

mortality) produced from the reports of male and female respondents indicate broadly the same level of mortality. However, the estimates of the mortality of men (i.e. the brothers' mortality) based on the reports of male respondents indicate much lower mortality than the estimates for men based on data supplied by women. While the latter estimates suggest that the mortality of men and women in early adulthood is broadly similar, the estimates for men produced from data supplied by male respondents suggest that young men have much lower mortality than young women in Bangladesh.

Clearly these estimates are of rather poor quality. They may severely underestimate the mortality of young adults in Bangladesh. One quite plausible explanation of the apparent discrepancies between the different series of estimates is that men are more likely to omit dead siblings from their reports than women, but that downward bias resulting from clustering of mortality at the sibship level is more severe in the estimates based on the reports made by same-sex siblings than the estimates based on the reports made by opposite sex siblings. The estimates based on men's reports about their brothers are particularly low as they suffer severely from both biases. In contrast, the two series for women appear fairly consistent because each is severely affected by one bias but less affected by the other. The implication of this pattern of errors, if it is indeed the explanation of the differences between the series, is that both sets of estimates for women probably underestimate their mortality by more than the estimates for brothers based on the reports of women. Thus, the mortality of young women in Bangladesh may well remain higher than that of men of the same age.

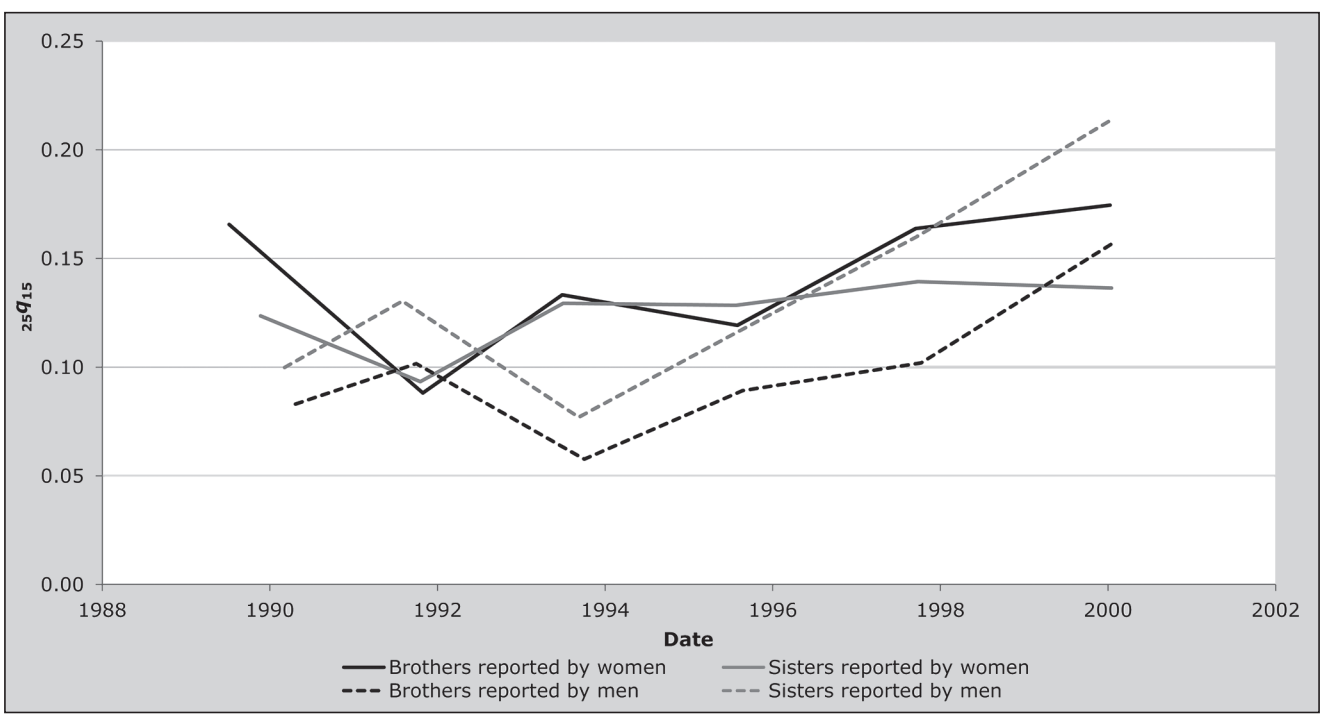
## DETAILED DESCRIPTION OF METHOD

### Introduction

The initial methods developed for estimating mortality from information on the survival of siblings were based on the idea that, on average, the ages of siblings are close to the age of a respondent. The proportion of a respondent's siblings who are still alive is, therefore, a good estimator of life table survivorship to the age of the respondent (Hill and Trussell 1977; UN Population Division 1983).

Unfortunately, field experience of this approach demonstrated that the quality of the data collected on

*FIGURE 23.1 Trends in the probability of dying between exact ages 15 and 40 estimated from adult siblings, Bangladesh, 2003 World Health Survey*



siblings was often low because siblings who died before or shortly after the respondent's own birth were often omitted by respondents, who may not know about them at all (Blacker and Brass 1983; Zaba 1986).

Interest in estimating mortality from data on siblings was revived by the development of the sisterhood method for measuring maternal mortality (Graham, Brass and Snow 1989). This requires data on how many sisters of the respondent survived to the age of 15, how many of them died thereafter, and whether sisters who died did so during pregnancy or within 6 to 8 weeks of the end of a pregnancy. Limiting the consideration of siblings to only those who survived to age 15 years excludes siblings who died while still young and, therefore, may have been unknown to or forgotten by the respondent. The responses supplied to the first two of these questions by respondents allow one to calculate the proportions still alive of sisters who survived to age 15 for each five-year age group. These proportions can be used to estimate the all-cause mortality of adult women. Comparable data on respondents' brothers can be used to estimate the mortality of men.

Thus, the only information required to apply the indirect adult sibling method is summary data on the proportion of adult sisters and brothers that are still alive among those who survived to age 15 tabulated by the age group of the respondents. As the siblings are about the same age on average as the respondents, for respondents aged $x$, these proportions approximately equal $l_x/l_{15}$. Because the relationship that exists between this measure and the proportion of siblings who survived to adulthood that are still alive is a close one, it can be estimated relatively precisely even in populations with unusual age patterns of mortality such as those experiencing severe HIV epidemics.

**Mathematical exposition**

Using the probability approach developed by Goodman, Keyfitz and Pullum (1974), Timæus, Zaba and Ali (2001) show that, in a stable population, the number of siblings ever born $y$ years before a respondent currently aged $a$ is given by ${}_aV_y$:

$$ {}_aV_y = \int_s^\omega e^{-r(z-a)} f(z) l(z) f(z-y) dz, y \geq 0 \qquad (1) $$

$$ {}_aV_y = \int_s^\omega e^{-r(z-a)} f(z) l(z-y) f(z-y) dz, y < 0 \qquad (2) $$

where Equation 1 gives the number of older siblings, Equation 2 the number of younger siblings, integration is over all ages at childbearing $s$ to $\omega$, and

$z$ = the age of the mother at the birth of the respondent

$f(z-y)$ = the probability of the respondents' mothers giving birth at age $z-y$ conditional on having given birth to the respondent at age $z$

$r$ = the growth rate in a stable population.

Note that in Equations 1 and 2, $f(x)$ is a *birth distribution*, which is to say the distribution of ages at giving birth of an individual woman, not the fertility distribution of the entire population.

The proportion of siblings still alive among those who lived to age 15 for respondents in a five-year age group, $x$ to $x+5$ is given by

$$ {}_5S_x^{15+} = \frac{\int_x^{x+5} l(a) \int_{15-a}^{\omega-s} {}_aV_y l(a+y) dy da}{l(15) \int_x^{x+5} l(a) \int_{15-a}^{\omega-s} {}_aV_y dy da}, x \geq 15. \qquad (3) $$

Calculation of the proportion of siblings alive for a given age of respondent requires a model of the birth distributions of individual women. Hill and Trussell (1977) assumed that all mothers experience the age-specific fertility rates of the general population. Thus, they could derive a sibling age distribution as a convolution of the fertility distribution. However, if women start childbearing at a wide range of ages, but compress all their childbearing into a small part of their fertile life span, as is typical in low fertility populations, one would expect the variance of the birth distribution to be considerably lower than the variance of the fertility schedule.

By contrast, when developing the sisterhood method, Graham, Brass, and Snow (1989) assumed that ${}_aV_y$, the distribution of age differences of siblings, could be represented by a normal distribution with a mean of zero and a variance of 80 years-squared. This assumption considerably simplifies the process of estimating the proportion of siblings who remain alive but is difficult to justify on theoretical grounds. In particular, the distribution of age differences between siblings would only be normal if the mother's birth distribution itself was normal. Using a normal distribution for the sibling age difference distribution constitutes a reasonable approximation if the birth distribution is peaked (i.e. $\sigma^2<35$), but is less satisfactory for representing sibling age difference distributions in the case of flatter birth distributions, such as occur in high- and medium-fertility populations.

A further issue arises in growing or shrinking populations. Goldman (1978) proved that, in a growing population, an individual selected at random from those whose mothers have completed childbearing has more younger siblings ever born than older ones. The opposite is true in a shrinking population. One can understand this intuitively by considering respondents currently aged 40, all of whose mothers have completed childbearing. In a growing stable population, relatively more of these respondents will have young mothers (say those currently aged less than 65 if they have survived) than in a stationary population because, at the time of their birth, there would have been more women aged less than 25 than in the corresponding stationary population. But, if the respondents are children of young mothers, they are more likely to have younger than older siblings because their mothers have more childbearing before them than behind them.

Thus, the distribution of sibling age differences is not symmetrical: its mean lies below zero in a growing population, while the opposite is true in a shrinking population. More precisely, if the variance of the underlying birth distribution is $\sigma^2$, then the mean of the sibling age distribution lies at approximately $-r\sigma^2$, where $r$ is the population growth rate. Thus, even if all women experienced the same age-specific fertility, the variance of the sibling age distribution in a growing population would still be slightly less than twice the variance of the fertility distribution and the distribution would be positively skewed. The opposite features characterize this distribution in shrinking populations.

In order to address these issues, Timæus, Zaba and Ali (2001) proposed a model of the sibling age differences that synthesized the two earlier approaches. On the basis of an investigation of the distributions of ages of older siblings reported in the birth histories collected by 12 nationally-representative surveys conducted as part of the World Fertility Survey (WFS), they concluded that the variances of birth distributions in the developing world range from about 45 to 110 years-squared. They then adapted the relational Gompertz model of fertility (Brass 1974, 1981) to represent these birth distributions, setting the $\beta$ parameter in their set of models to values that vary between 1 and 1.8 to produce distributions of the appropriate width (as $\beta$ increases the variance of the model distributions decreases). To allow for the absence of very short birth intervals in human populations, ${}_aV_0$ was set to 0 and ${}_aV_1$ and ${}_aV_{-1}$ to 40 per cent of the model values. The value of 40 per cent reproduces the average of the ratios ${}_aV_1 / {}_aV_2$ in the 12 WFS populations.

**Implementation of the method**

Although Equation 3 has to be evaluated numerically, in principle there is no reason it could not be solved directly for life table survivorship using the Excel *Solver* routine or a similar tool and a birth distribution, ${}_aV_y$ that is appropriate for the population under study. To arrive at a unique solution, an assumption still has to be made about the age pattern of mortality within adulthood such as which standard to adopt in a 1-parameter system of model life tables. In practice, estimates are usually produced using regression models that have been fitted to simulated data on the survival of siblings generated for populations with a wide range of age structures, birth distributions and mortality schedules (Timæus, Zaba and Ali 2001). The coefficients of these models are shown in Table 23.1.

***Mathematical exposition – time location of the estimates***

Time location methods aim to estimate the time $T$ at which the cohort measures of survival that produced the proportion of relatives surviving, ${}_a p_y^c$, equalled the equivalent period measures, ${}_a p_y(T)$, so that

$$S(a) = \int_s^\omega {}_a v_y \, {}_a p_y(T) dy$$

where:

$${}_a v_y = \frac{{}_aV_y}{\int_s^\omega {}_aV_y dy}$$

If we denote the mean time since death of those dying between $y$ and $y+a$ by ${}_a g_y$, by definition

$${}_a g_y = \frac{\int_y^{y+a} \left(z - y - a\right) \mu(z) l(z) dz}{\int_y^{y+a} \mu(z) l(z) dz} \qquad (4)$$

where $\mu(z)l(z)$ is the life table deaths at age $z$. Brass and Bamgboye (1981) show that, if mortality schedules conform to a system of 1-parameter relational logit model life tables and the trend in adult mortality is assumed to be linear in $\alpha$, the parameter of that relational system of models, the time at which the cohort survivorship of adults equals period survivorship is a weighted average of the times since death of the respondents' relatives:

$$T = \frac{\int_{x}^{x+5} l(a) \int_{s}^{\omega} {}_{a}v_{y}\, l(y+a,T)\left(1 - {}_{a}p_{y}(T)\right) {}_{a}g_{y}\, dy\, da}{\int_{x}^{x+5} l(a) \int_{s}^{\omega} {}_{a}v_{y}\, l(y+a,T)\left(1 - {}_{a}p_{y}(T)\right) dy\, da}. \quad (5)$$

This time depends on the level of mortality and the ages of the relatives but is independent of the rate of change in $\alpha$. Although Brass and Bamgboye's derivation of Equation 3 takes advantage of a relationship between changes in mortality with age and with time that is specific to a relational logit system of life tables, it is possible to arrive at similar formulae for $T$ on the basis of other reasonable assumptions about the trend in mortality with time by age (Palloni, Massagli and Marcotte 1984).

Equation 5 can be evaluated numerically, using values for ${}_{a}v_{y}$ and for the life table measures chosen on the basis of observed data. To develop a straightforward procedure for estimating $T$ from observed characteristics of a population, a much simpler relationship must be assumed. Brass and Bamgboye (1981) argue that the change in $T$ with $a$ over limited age ranges is sufficiently close to linear that all respondents in a five-year age group can be treated as of the central age $N$. Second, they argue, at the ages and levels of mortality at which indirect methods are used to estimate adult mortality, the force of mortality increases approximately exponentially with age. As a consequence, for such applications, variation in ${}_{a}g_{y}$ with $y$ is slight. Therefore, the weighting factors for ${}_{a}g_{y}$ in Equation 5 have little effect and all adult relatives can be treated as entering exposure at their mean age of entry, $M$. To a satisfactory approximation,

$$T = {}_{N}g_{M}.$$

If survivorship in adulthood fell linearly with age, so that the same number of deaths occurred at every age, then ${}_{N}g_{M}$ would be $N/2$ whatever the value of $M$. In less extreme life tables, mortality rises with age more rapidly than this, and the deaths of the relatives are concentrated at older ages and, therefore, in the recent portion of the $N$-year period. This means that the time location of the estimates is closer to the survey date than $N/2$. By substituting $e^{kz}\mu(y)l(y)$ for $\mu(z)l(z)$ in Equation 4 and expanding the right-hand side in powers of $N$, Brass and Bamgboye (1981) demonstrate that the appropriate adjustment to ${}_{N}g_{M}$ is a function of both $k$ and $N$:

$${}_{N}g_{M} \approx \frac{N}{2}\left(1 - k\frac{N}{6}\right). \quad (6)$$

Brass and Bamgboye (1981) also demonstrate that the assumption that mortality increases exponentially with age implies that in a relational logit life table system

$$\frac{e^{kN}}{e^{k_s N}} = \frac{\left({}_{N}p_{M}\right)^2}{\left(l^s(M+N)/l^s(M)\right)^2}.$$

Solving for $kN$ and substituting this expression into Equation 6 yields an estimate of ${}_{N}g_{M}$, and therefore of $T$, of:

$$T = \frac{N}{2}\left(1 - \frac{\ln\left({}_{N}p_{M}\right)}{3} + \frac{1}{3}\ln\left(\frac{l^s(M+N)}{l^s(M)}\right) - \frac{k_s N}{6}\right).$$

Thus, in this formulation, the time references of measures of conditional survivorship obtained from data on adult relatives are estimated as half the duration of exposure, $N$, reduced by a factor that depends on the level of conditional survivorship relative to a standard life table.

Having arrived at this expression for $T$ on theoretical grounds, Brass (1985) approximates ${}_{N}p_{M}$ by ${}_{5}S_{x}$ and adopts as his standard life table one in which $l^s(x)$ is linear over the adult ages and is taken as $(1-x/80)/2$. As $l^s(x)$ is linear, $T = \frac{1}{2}N$ and $k_s$ becomes 0. Thus, $T$ is estimated from observed data using

$$T = \frac{N}{2}\left(1 - \frac{\ln\left({}_{5}S_{x}\right)}{3} + \frac{1}{3}\ln\left(\frac{80-M-N}{80-M}\right)\right). \quad (7)$$

In the adult sibling method, $M$, the age at which exposure begins, is exactly 15 years for every sibling. The asymmetry of the sibling age difference distribution means that, in a growing population, the siblings are on average slightly younger than the respondents. This age difference varies between about zero and 1.75 years in those populations in which one is likely to want to apply the method. One can use a central value of 0.8 years in all applications. Thus, the duration of exposure, $N$, becomes $(n-2.5-0.8)-15$, where $n$ is still the upper limit of the age group of respondents. Because $M$ is fixed at 15 years, Equation 7 can be simplified for each age group to a linear equation of the form (Timæus, Zaba and Ali 2001)

$$T = a(n) - b(n)\ln\left({}_{5}S^{15+}_{n-5}\right).$$

## PERFORMANCE IN POPULATIONS WITH GENERALIZED HIV EPIDEMICS

The HIV epidemic poses two problems for indirect methods of estimating mortality based on the survival of relatives. First, both the sexual and vertical routes of transmission produce significant selection biases in data collected in surveys on the survival of relatives. Second, the incidence of HIV infection is concentrated among young adults. Thus, populations with significant AIDS mortality have very different age patterns of mortality both from other populations and from the model life tables used to derive coefficients for converting data on survival of relatives into measures of life table survivorship.

A major advantage of sibling methods of measuring adult mortality, compared with the orphanhood method, is that they are free of selection biases arising from direct transmission of the virus. Some residual bias due to clustering of AIDS mortality within sibships will remain. In particular, the risk of HIV infection tends to vary markedly between localities and siblings often live close to each other. The impact of this, however, will be relatively small compared with the biases that affect data that parents have supplied about their children or vice versa.

Bias in the regression coefficients used to estimate life table survivorship remains more of a problem. With respect to Equation 3, it is the change in the age pattern of mortality experienced by the siblings as a result of AIDS that is of concern, not the impact of the epidemic on the sibling age difference distribution, as the main factor shaping this distribution is the age pattern of childbearing rather than mortality or age structure.

Timæus, Zaba and Ali (2001) assess the sensitivity of the adult sibling method estimates to these problems using a combination of empirical and simulated data. They find that even in the presence of the unusual age pattern of mortality found in populations with high AIDS mortality, the adult sibling method produces estimates of survivorship that are close to the actual values. The estimates based on data on respondents aged 20–24 years and more than 40 years are extremely accurate, while those based on data for respondents aged 25 to 39 years slightly overestimate survivorship. This is because the regression coefficients fail to allow for the concentration of AIDS deaths in this age range.

To use sibling estimates of adult survivorship to monitor mortality trends, it is necessary to fit a model life table to the estimates for specific age ranges and use it to extrapolate to an index referring to a common range of ages. Somewhat surprisingly, if one converts the entire series of estimates to measures of survivorship from 15 to 50 years, ${}_{35}p_{15}$, these remain fairly accurate. Those obtained from respondents aged 25 to 34 are more accurate than the estimates of $l_n/l_{15}$ on which they are based. Errors due to the failure to allow for the impact of AIDS on the mortality schedule in first calculating the coefficients and then extrapolating to a common measure of survivorship largely cancel out. This finding is robust to variation in background mortality and choice of a mortality standard. Thus, estimates of ${}_{35}p_{15}$ obtained from the adult sibling method probably represent relatively robust indices for the monitoring of mortality trends as the AIDS epidemic develops. As with other indirect methods, if successive sets of data are collected for the same population, checks on the consistency of the estimates for periods when they overlap provide a powerful indication of the accuracy of the results.

## EXTENSIONS AND VARIANTS OF THE METHOD

Most surveys that have collected the information required to estimate all-cause mortality of adults from data on adult siblings have also asked whether dead sisters died while pregnant or shortly after giving birth. Together these data provide the basis for applying the sisterhood method for estimating maternal mortality (Graham, Brass and Snow 1989). It is also possible to calculate direct sibling estimates of adult mortality from the detailed sibling histories collected in many Demographic and Health Surveys and some other inquiries.

## FURTHER READING AND REFERENCES

The adult sibling method is not discussed in the classic manuals on indirect estimation (Sloggett, Brass, Eldridge *et al.* 1994; UN Population Division 1983) but is described in the United Nations manual on estimating adult mortality (UN Population Division 2002). The key reference explaining the theoretical basis of the adult sibling method and the development of the regression coefficients for conversion of proportions of surviving siblings into life table indices is Timæus, Zaba and Ali (2001). This article surveys earlier contributions to the literature.


Blacker JGC and W Brass. 1983. "Experience of retrospective enquiries to determine vital rates," in Moss, L and H Goldstein (eds). *The Recall Method in Social Surveys.* London: University of London Institute of Education, pp. 48–61.

Brass W. 1974. "Perspectives in population prediction: illustrated by the statistics of England and Wales", *Journal of the Royal Statistical Society* **A137**(4):532–583.

Brass W. 1981. "The use of the Gompertz relational model to estimate fertility," Paper presented at International Population Conference, Manila, 1981. Liège. International Union for the Scientific Study of Population. Vol. 3:345–362.

Brass W. 1985. *Advances in Methods for Estimating Fertility and Mortality from Limited and Defective Data.* London: London School of Hygiene & Tropical Medicine.

Brass W and EA Bamgboye. 1981. *The Time Location of Reports of Survivorship: Estimates for Maternal and Paternal Orphanhood and the Ever-widowed*. London: London School of Hygiene & Tropical Medicine.

Coale AJ, P Demeny and B Vaughan. 1983. *Regional Model Life Tables and Stable Populations*. London: Academic Press.

Gakidou E and G King. 2006. "Death by survey: estimating adult mortality without selection bias from sibling survival data", *Demography* **43**(3):569–585. doi: 10.1353/dem.2006.0024

Goldman N. 1978. "Estimating the intrinsic rate of increase of a population from the average numbers of younger and older sisters", *Demography* **15**(4):499–521. doi: 10.2307/2061202

Goodman LA, N Keyfitz and TW Pullum. 1974. "Family formation and the frequency of various kinship relationships", *Theoretical Population Biology* **5**(1):1–27. doi: 10.1016/0040-5809(74)90049-5

Graham W, W Brass and RW Snow. 1989. "Estimating maternal mortality: The sisterhood method", *Studies in Family Planning* **20**(3):125–135. doi: 10.2307/1966567

Hill K and J Trussell. 1977. "Further developments in indirect mortality estimation", *Population Studies* **31**(2):313–334. doi: 10.2307/2173920

Masquelier B. 2013. "Adult mortality from sibling survival data: A reappraisal of selection biases?", *Demography* **50**(1):207–228. doi: 10.1007/s13524-012-0149-1

Obermeyer Z, JK Rajaratnam, CH Park, E Gakidou *et al*. 2010. "Measuring adult mortality using sibling survival: a new analytical method and new results for 44 countries, 1974–2006", *PLoS Medicine* **7**(4):e1000260. doi: 10.1371/journal.pmed.1000260

Palloni A, M Massagli and J Marcotte. 1984. "Estimating adult mortality with maternal orphanhood data: analysis of sensitivity of the techniques", *Population Studies* **38**(2):255–279. doi: 10.1080/00324728.1984.10410289

Sloggett A, W Brass, SM Eldridge, IM Timæus, P Ward and B Zaba. 1994. *Estimation of Demographic Parameters from Census Data*. Tokyo, Japan: United Nations Statistical Institute for Asia and the Pacific.

Timæus IM, B Zaba and M Ali. 2001. "Estimation of adult mortality from data on adult siblings," in Zaba, B and J Blacker (eds). *Brass Tacks: Essays in Medical Demography*. London: Athlone, pp. 43–66.

Trussell J and G Rodriguez. 1990. "A note on the sisterhood estimator of maternal mortality", *Studies in Family Planning* **21**(6):344–346. doi: 10.2307/1966923

UN Population Division. 1982. *Model Life Tables for Developing Countries*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/77. http://www.un.org/esa/population/techcoop/DemMod/model_lifetabs/model_lifetabs.html

UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation.* New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

UN Population Division. 2002. *Methods for Estimating Adult Mortality.* New York: United Nations, Department of Economic and Social Affairs, ESA/P/WP.175. http://www.un.org/esa/population/techcoop/DemEst/methods_adultmort/methods_adultmort.html

Zaba B. 1986. *Measurement of Emigration using Indirect Techniques: Manual for the Collection and Analysis of Data on Residence of Relatives*. Liège: Ordina.

Zaba B and PH David. 1996. "Fertility and the distribution of child mortality risk among women", *Population Studies* **50**(2):263–278. doi: http://dx.doi.org/10.1080/0032472031000149346

# Multiple Census Methods

# Chapter 24 The generalized growth balance method

*Rob Dorrington*

## DESCRIPTION OF METHOD

Hill (1987) generalized the Brass Growth Balance method for estimating the completeness of the reporting of deaths relative to an estimate of the population on the assumption that the population was demographically stable, to non-stable populations closed to migration. This generalization can be used where one has data on the numbers by age group from two censuses and an estimate of the number of deaths by age group between the dates of the two censuses. With the additional information from two censuses it is possible to estimate age-specific growth rates in place of the constant growth rate implied by the assumption of stability. The method still assumes, however, that the proportion of deaths reported and the completeness of the census counts is the same at all adult ages and that, apart from this, the data are accurately reported. Moreover, in its common formulation it assumes that the population is closed to migration, although the method can be adapted to accommodate migration if data are available.

In all closed populations, $r(x+) = b(x+) - d(x+)$, where the partial 'birth' rate, $b(x+)$, is defined as the rate at which people turn age $x$ in the population aged $x$ and older and the partial death rate, $d(x+)$, is the rate of mortality of people aged $x$ and older. If, in this population, the deaths are under-reported to the same extent at each age then $b(x+) - r(x+) = d^r(x+)/c$, where $d^r(x+)$ is the recorded death rate for ages $x$ and older and $c$ is the proportion of deaths that are reported. In practice, the count of the census populations from which $r(x+)$ is estimated may not be complete but the assumption that the undercount is the same at each age makes it possible to solve for $c$ from the slope of the line fitted to the $b(x+) - r(x+)$ and $d^r(x+)$ data points. Mortality rates can then be estimated by dividing the numbers of deaths reported in each age group by $c$ and dividing these numbers by an estimate of the population exposed to risk, to estimate the partial birth, growth and death rates. Moreover, as a by-product of the procedure, the less complete census counts can be adjusted to be mutually consistent, although not necessarily accurate.

## DATA REQUIREMENTS AND ASSUMPTIONS

### Tabulations of data required

- Number of women (men), by five-year age group, and for open age interval $A+$ (with $A$ as high as possible), at two points in time, typically from the results of two censuses. (See the caveat below concerning the use of surveys instead of censuses.)
- Number of deaths of women (men), by five-year age group, and for open age interval $A+$, over the period between the two censuses or surveys.

### Important assumptions

- The coverage of each census is the same for all ages.
- The completeness of reporting of deaths is the same for all ages above a minimum age (usually age 5 or 15).
- The population is closed to migration. Although the method can be adapted to allow for migration, accurate enough estimates of the net numbers of migrants to do so seldom exist. For national populations, net migration is often low enough to ignore, but for situations where migration is significant one needs to take this into account when interpreting results and deciding on an estimate of completeness.

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Before applying this method, you should investigate the quality of the data in at least the following dimensions:

- age structure of the population;
- sex structure of the population;
- age structure of the deaths; and
- sex structure of the deaths.

If the reported deaths are for a period other than that between the censuses, the numbers that would have been reported in the intercensal period need to be estimated. If one has annual vital registration data, this adjustment involves apportioning deaths in the first and last year of the period. If one has deaths reported by households the year before the dates of each of the first and second censuses, one has to estimate the numbers of deaths by interpolating between these estimates for the intercensal period (using the *Estimating deaths.xlsx* spreadsheet).

## CAVEATS AND WARNINGS

In applying this method, analysts must take particular care with the following.

The interpretation and estimation processes need to take into account the source of death data (vital registration, reported by households in censuses, or recorded in hospitals) as explained below. However, the biases associated with the source of death data tend to have less impact on the estimate of completeness from the Growth Balance method than on the Synthetic Extinct Generations method.

- If applying the method to sub-national geographic areas, the issue of migration typically becomes a greater concern.
- Deciding the age range to be used to fit the straight line to the partial birth and death rates and hence estimate completeness. Issues here are: the best age to choose for the open interval if there is evidence of age exaggeration; how to accommodate data points that rise above the line at the older ages because of falling completeness possibly due to retirement-associated migration from urban to rural areas where registration is less complete; and whether to exclude ages less than either 30 or 35 because of the impact of migration which has not been allowed for specifically.
- If completeness appears to be less than 60 per cent then the uncertainty is large and this should be taken into account when interpreting the results.
- It is tempting in a situation in which census data on the age distribution of the population and household deaths are available for only one census to use in this method sample survey data on the age distribution of the population at some earlier or later date. However, for reasons that have not been adequately researched, such a combination of data sources rarely gives satisfactory results.

## APPLICATION OF METHOD

Although technically one could apply this method to data in single year age categories, the data one typically works with are subject to age misstatement, so in practice one usually works with data grouped into five-year age groups. For convenience, since most data are published in this format, the spreadsheet is set up to work with data in the standard five-year groupings. However, as Blacker (1988) has shown, if this grouping fails to remove the effect of digit preference, the method should be adapted to work with an alternative five-year grouping of ages centred on, rather than starting with, ages at which heaping occurs.

### *Step 1: If not readily available, estimate the number of deaths reported in the period between the dates of the two estimates of the population*

In the case where one has annual vital registration data, this adjustment involves apportioning deaths in the first and last year of the period to the parts of the year before and after the mean dates of fieldwork of the two inquiries. Unless the age pattern of deaths is changing very rapidly, this approximation will have no effect on the results.

If one lacks data on the number of deaths between the two inquiries but this interval falls between two periods for which one does have such estimates (for example, because each inquiry included a question about deaths in the household during the previous year), one can use the *Estimating deaths* spreadsheet. This spreadsheet estimates the number of deaths between two points in time given estimates of deaths over two other periods. To use this spreadsheet, you need the number of deaths divided into five-year age groups for two periods (periods 1 and 2), the start and end dates for each of these periods, and the start date and end date of the period for which one wishes to estimate the number of deaths.

### *Step 2: Cumulate population, deaths and migrants downwards*

To estimate partial birth, death (and migration) rates one needs to cumulate the numbers in the population, and the number of deaths (and the net number of migrants) for ages $x$ and older. Thus, in the case of the population the following equation is used:

$$N(x+) = \sum_{y=x}^{A-5} {}_5N_y + {}_\infty N_A$$

where $A$ is the age at the start of the open age interval.

Analogous equations are used to calculate the number of deaths aged $x$ and older, $D(x+)$. In the case where these are available (unlikely though this may be) analogous equations can be used to calculate the net number of migrants aged $x$ and older, $NM(x+)$). Where the numbers of migrants is not known this column is set to zero (or left blank) and the method is applied taking into account this omission, as described below.

***Step 3: Calculate the person-years of life lived, PYL(x+)***

In order to estimate partial birth and death rates (and if one has data on the net numbers of migrants, migration rates) one needs to estimate the person-years of exposure. This is estimated using the following formula:

$$PYL(x+) = (t_2 - t_1)\left( {}_{\infty}N_x(t_1) \times {}_{\infty}N_x(t_2)\right)^{1/2}$$

where $t_1$ is the time of the first census, and $t_2$ the time of the second census.

***Step 4: Calculate the number of people who turned x in the population, N(x)***

The number of people who turned $x$ (i.e. were 'born' into the open age interval $x+$) in the population is estimated as the geometric mean of the numbers in a cohort at times $t_1$ and $t_2$ divided by 5, multiplied by the length of the period between the censuses, in years, using the following formula:

$$N(x) = \frac{t}{5}\left( {}_5N_{x-5}(t_1) \times {}_5N_x(t_2)\right)^{\frac{1}{2}}.$$

***Step 5: Calculate partial birth and death rates, b(x+) and d(x+), and partial growth rate r(x+) corrected for migration, i(x+)***

The partial birth and death rates are estimated using the following formulae:

$$b(x+) = \frac{N(x)}{PYL(x+)}$$

$$d(x+) = \frac{D(x+)}{PYL(x+)}$$

while the partial growth rate less the partial migration rate is calculated using the following formula:

$$r(x+) - nm(x+) = \frac{{}_{\infty}N_x(t_2) - {}_{\infty}N_x(t_1) - NM(x+)}{PYL(x+)}.$$

***Step 6: Plot graph of b(x+)–r(x+)+nm(x+) against d(x+), and examine to decide on the range over which the line should be fitted***

Start by setting the lower age to 5 and the upper age to $A$–1, where $A$ is the age at the start of the open interval of the data. Inspect the diagnostic plots and decide on the age interval over which the line is to be fitted. If there is greater age exaggeration in ages at death than in ages of the living the points plotted to the right (older ages) will fall progressively below the line with age. This indicates that a lower maximum age is called for – stepping down in five-year steps until the effect is removed. Also, if the absolute value of the residuals of the end points are too large (e.g. exceed 0.01) then the maximum age should be lowered to prevent these points unduly influencing the slope of the line. If the age exaggeration is the same in the population and the deaths then this will have no effect on the slope and hence the estimate of completeness of reporting, but the age-specific death rates will be biased downward for these ages.

If the points plotted at the younger ages (left-hand side), particularly ages 15 to 30, deviate noticeably from the straight line and one has not included any data on migration, this probably indicates that there is significant migration (unless there is age differential under-enumeration). One should thus increase the lower age of the age interval used to fit the line to age 30 or 35, depending on which produces the most sensible fit to the data.

***Step 7: Fit line and estimate completeness, c***

In order to estimate the completeness of reporting of deaths relative to the population, one starts by plotting $b(x+)$–$r(x+)+i(x+)$ against $d(x+)$ and estimating the coefficients of the straight line fitted to these points, using orthogonal regression, as follows:

$$b = \frac{\sigma_y}{\sigma_x} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \mu_y\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \mu_x\right)^2}$$

and

$$a = \mu_y - b\mu_x$$

where $b$ is the slope of the line and $a$ the intercept, the $y_i$ represent the $b(x+)$–$r(x+)+i(x+)$, the $x_i$ represent the $d(x+)$ and $\mu_y$ and $\mu_x$ represent the means of the two series, respectively.

After fitting the straight line to all the points, one inspects the plotted points relative to the line and the residuals in order to decide on the best range of ages to use to determine the completeness of reporting of deaths. How one decides this is discussed in more detail below but any residuals greater than 1 per cent in absolute value should be excluded. A line is then fitted to these points, from which new values of $a$ and $b$ are determined. As a general rule, it is not recommended to truncate at an age ending in zero in a population with significant digital preference.

The completeness of reporting of deaths, $c$, is derived from the values of $a$ and $b$ as follows. Since

$$a = \frac{\ln\left(k_1/k_2\right)}{t_2 - t_1} \text{ and } b = \left(\frac{k_1 k_2}{c}\right),$$

$$\frac{k_1}{k_2} = e^{a(t_2 - t_1)}$$

and $c$ is estimated by assuming the larger of $k_1$ and $k_2$ =1.

Thus if $\frac{k_1}{k_2} < 1$, assume that $k_2 = 1$ and hence

$$k_1 = e^{a(t_2 - t_1)} \text{ and } c = \frac{e^{a(t_2 - t_1)}}{b}$$

and if $\frac{k_1}{k_2} > 1$, assume that $k_1 = 1$ and hence

$$k_2 = e^{-a(t_2 - t_1)} \text{ and } c = \frac{e^{-a(t_2 - t_1)}}{b}.$$

***Step 8: Estimate mortality rates adjusted for incompleteness of reporting of deaths***

In order to compute mortality rates one needs first to correct the census population for relative under-enumeration. This is achieved by dividing the numbers from the first census by $k_1$ and the numbers from the second census by $k_2$.

Next one needs to adjust the number of deaths for incompleteness by dividing the reported number of deaths by the estimate of completeness, $c$.

The adjusted person-years of exposure, $PYL^a(x,5)$, are estimated in the same way as before but using the population corrected for under-enumeration as follows:

$$PYL^a(x,5) = \left(t_2 - t_1\right)\left({}_5N_x(t_1)/k_1 \times {}_5N_x(t_2)/k_2\right)^{1/2}.$$

Next one needs to adjust the number of deaths for incompleteness by dividing the reported number of deaths by the estimate of completeness, $c$, and dividing this by $PYL^a(x,5)$ to produce mortality rates adjusted for the incompleteness of the reporting of deaths as follows:

$${}_5m_x = \frac{{}_5D_x/c}{PYL^a(x,5)}.$$

Note that technically one could drop the $k_1/k_2$ adjustment and still get the same estimates of the mortality rates (since the same adjustment is made to both the numerator and the denominator). However, in that case the estimate of completeness is relative to the average of the census populations ignoring the fact that one is undercounted relative to the other.

***Step 9: Smooth using relational logit model life table***

Because the age-specific rates can be quite erratic they need to be graduated (smoothed). This can be achieved by fitting a Brass relational logit function to a sex-specific standard life table which is considered to have the same shape as that generated by the mortality rates of the population being investigated.

The workbook (see website) contains a spreadsheet that allows one to produce a smooth set of mortality rates by using a relational logit model fitted to the life table generated by the adjusted mortality rates. The user can choose between the standard from the General family of United Nations model life tables or one from any of the four families of Princeton model life tables. The logit transforms of these tables together with a model life table of a population experiencing an AIDS epidemic (Timæus 2004) appear in the *Models* spreadsheet. This spreadsheet also allows the user to input logit transforms of an alternative life table if there is reason to assume that it has a similar pattern of adult mortality to that of the population being studied.

In order to fit the model, probabilities of people aged $x$ dying in the next 5 years, ${}_5q_x$, are estimated from the adjusted rates of mortality as follows:

$${}_5q_x = \frac{5\,{}_5m_x}{1 + 2.5\,{}_5m_x}.$$

From this the life table with a radix of $l_5 = 1$ is calculated as follows:

$$l_{x+5} = l_x\left(1 - {}_5q_x\right).$$

The coefficients, $\alpha$ and $\beta$ are determined by fitting the relational logit model as follows:

$$\gamma_x = \alpha + \beta\gamma_x^s$$

where

$$\gamma_x = 0.5\ln\left(\frac{1-l_x}{l_x}\right)$$

and superscript '*s*' designates values based on a standard life table.

The fitted life table is then generated from the standard life table using the coefficients $\alpha$ and $\beta$ as follows:

$$\gamma_x^{fitted} = \alpha + \beta\gamma_x^s$$

and

$$l_x^{fitted} = \frac{1}{\exp\left(1+2\gamma_x^{fitted}\right)}.$$

The smoothed mortality rates are derived from this life table as follows:

$${}_5m_x^{fitted} = \frac{l_x^{fitted} - l_{x+5}^{fitted}}{T_x - T_{x+5}}$$

and

$${}_\infty m_x^{fitted} = \frac{l_x^{fitted}}{T_x}$$

where

$$T_x = \sum_{x=x,5}^{\omega} \frac{5}{2}\left(l_x^{fitted} + l_{x+5}^{fitted}\right)$$

i.e.

$$T_x = T_{x+5} + \frac{5}{2}\left(l_x^{fitted} + l_{x+5}^{fitted}\right)$$

and $\omega$ is the age above which the life table has no more survivors.

## WORKED EXAMPLE

This example uses data on the numbers of males in the population from the South African Census in 2001 and the Community Survey in 2007, on number of deaths from vital registration for the years 2001 to 2007, and on the net number of migrants estimated from the change in foreign-born counted in the two surveys, less an estimate of the number of South Africans who emigrated between the two surveys. The example appears in the *GGB_South Africa_males* workbook.

### *Step 1: If not readily available, estimate the number of deaths reported in the period between the dates of the two estimates of the population*

The registered deaths for the years 2001 to 2007 for South African males are given in Table 24.1.

*TABLE 24.1 Calculation of deaths between census dates, South African males, 2001–2007*

| Age | 2001 | 2002–2006 | 2007 | Total between censuses |
|---|---|---|---|---|
| 0–4 | 29,005 | 186,346 | 40,314 | 197,912 |
| 5–9 | 2,118 | 14,733 | 2,854 | 15,566 |
| 10–14 | 1,745 | 10,535 | 2,233 | 11,207 |
| 15–19 | 4,470 | 23,857 | 4,860 | 25,473 |
| 20–24 | 8,931 | 51,588 | 10,875 | 54,960 |
| 25–29 | 16,834 | 96,705 | 18,405 | 102,802 |
| 30–34 | 20,892 | 137,355 | 28,245 | 145,588 |
| 35–39 | 21,068 | 137,502 | 29,258 | 145,900 |
| 40–44 | 19,322 | 128,217 | 26,973 | 135,936 |
| 45–49 | 17,881 | 113,891 | 24,761 | 121,010 |
| 50–54 | 16,883 | 104,508 | 22,790 | 111,157 |
| 55–59 | 14,544 | 90,919 | 21,317 | 96,854 |
| 60–64 | 15,097 | 84,351 | 17,410 | 89,930 |
| 65–69 | 13,011 | 77,680 | 17,878 | 82,843 |
| 70–74 | 14,035 | 68,147 | 13,771 | 73,036 |
| 75–79 | 10,846 | 59,859 | 12,534 | 63,871 |
| 80–84 | 9,161 | 44,986 | 8,872 | 48,163 |
| 85+ | 7,602 | 43,233 | 10,009 | 46,196 |

The reference time for the Census in 2001 was midnight between 9 and 10 October 2001. The Community Survey took place over a number of weeks in February so we can assume a reference time of midnight between 14 and 15 February 2007. Thus, if we assume deaths occur uniformly over the respective calendar years, we can apportion the deaths in 2001 and in 2007 and add these to the total for the years 2002 to 2006 to get the total number of deaths between the two estimates of the population. For example, for the age group 20–24 the number is calculated as follows:

$$\frac{(22+30+31)}{365}8931 + 51588 + \frac{(31+14)}{365}10875 = 54960.$$

### *Step 2: Cumulate population, deaths and migrants downwards*

One accumulates the numbers in the population, deaths and migrants from the oldest age downwards (Table 24.2).

*TABLE 24.2 Calculation of the cumulated populations, deaths and migrants, South African males, 2001–2007*

| Age | ${}_5N_x(t_1)$ | ${}_5N_x(t_2)$ | ${}_5D_x$ | ${}_5NM_x$ | *P1*(*x*+) | *P2*(*x*+) | *D*(*x*+) | *NM*(*x*+) |
|---|---|---|---|---|---|---|---|---|
| 0 | 2,223,006 | 2,505,744 | 197,912 | 10,605 | 21,434,045 | 23,348,679 | 1,568,404 | 128,946 |
| 5 | 2,425,066 | 2,560,642 | 15,566 | 2,848 | 19,211,039 | 20,842,935 | 1,370,492 | 118,341 |
| 10 | 2,518,985 | 2,452,339 | 11,207 | 5,153 | 16,785,973 | 18,282,293 | 1,354,926 | 115,492 |
| 15 | 2,453,156 | 2,553,293 | 25,473 | 16,574 | 14,266,988 | 15,829,955 | 1,343,719 | 110,339 |
| 20 | 2,099,417 | 2,362,519 | 54,960 | 14,803 | 11,813,832 | 13,276,662 | 1,318,246 | 93,766 |
| 25 | 1,899,275 | 2,033,165 | 102,802 | 4,714 | 9,714,415 | 10,914,143 | 1,263,286 | 78,963 |
| 30 | 1,594,624 | 1,875,483 | 145,588 | 13,331 | 7,815,140 | 8,880,977 | 1,160,484 | 74,249 |
| 35 | 1,441,657 | 1,548,185 | 145,900 | 9,693 | 6,220,516 | 7,005,495 | 1,014,896 | 60,918 |
| 40 | 1,233,813 | 1,306,900 | 135,936 | 7,464 | 4,778,859 | 5,457,310 | 868,996 | 51,225 |
| 45 | 967,744 | 1,104,294 | 121,010 | 8,719 | 3,545,046 | 4,150,410 | 733,060 | 43,761 |
| 50 | 769,627 | 888,042 | 111,157 | 9,413 | 2,577,302 | 3,046,116 | 612,050 | 35,042 |
| 55 | 552,402 | 708,812 | 96,854 | 4,640 | 1,807,675 | 2,158,074 | 500,893 | 25,629 |
| 60 | 444,592 | 491,871 | 89,930 | 5,081 | 1,255,273 | 1,449,261 | 404,039 | 20,989 |
| 65 | 304,835 | 394,305 | 82,843 | 4,922 | 810,681 | 957,391 | 314,108 | 15,908 |
| 70 | 232,604 | 241,976 | 73,036 | 4,334 | 505,846 | 563,086 | 231,266 | 10,986 |
| 75 | 136,466 | 163,112 | 63,871 | 2,980 | 273,242 | 321,110 | 158,229 | 6,652 |
| 80 | 90,856 | 87,698 | 48,163 | 1,662 | 136,776 | 157,998 | 94,359 | 3,672 |
| 85 | 45,920 | 70,299 | 46,196 | 2,009 | 45,920 | 70,299 | 46,196 | 2,009 |

***Step 3: Calculate the person-years of life lived, PYL(x+)***

Calculating person-years lived requires an estimate of the time between the two counts. This has been calculated using the YEARFRAC function in Excel on the basis of the date of the day following the time reference for the censuses. Counting days and dividing by 365 produces a slightly different estimate (5.3507 years) but has a negligible impact on the estimate of completeness.

The person-years of life lived is given in column 2 of Table 24.3 and is calculated from the numbers of the cumulated population in columns 2 and 3 of Table 24.2. For age 20, for example, as follows:

$$PYL(20+) = 5.3541\left(2099417 \times 2362519\right)^{1/2} = 67053861.$$

***Step 4: Calculate the number of people who turned x in the population, N(x)***

The numbers of people who turned *x* are shown in the third column of Table 24.3. For example, the number who turned 20 is estimated from the population numbers in columns 2 and 3 of Table 24.1 as follows:

$$N(20) = \frac{5.3541}{5}\left(2518985 \times 2553293\right)^{\frac{1}{2}} = 25778889.$$

***Step 5: Calculate partial birth and death rates, b(x+) and d(x+), and partial growth rate r(x+) corrected for migration, i(x+)***

The partial birth and death rates are shown in columns 4 and 6 of Table 24.3. The partial birth and death rates are calculated from the partial births (column 3 of Table 24.3) and the partial deaths (column 8 of Table 24.2) as follows for age 20, for example:

$$b(20+) = \frac{2577889}{67053861} = 0.03845$$

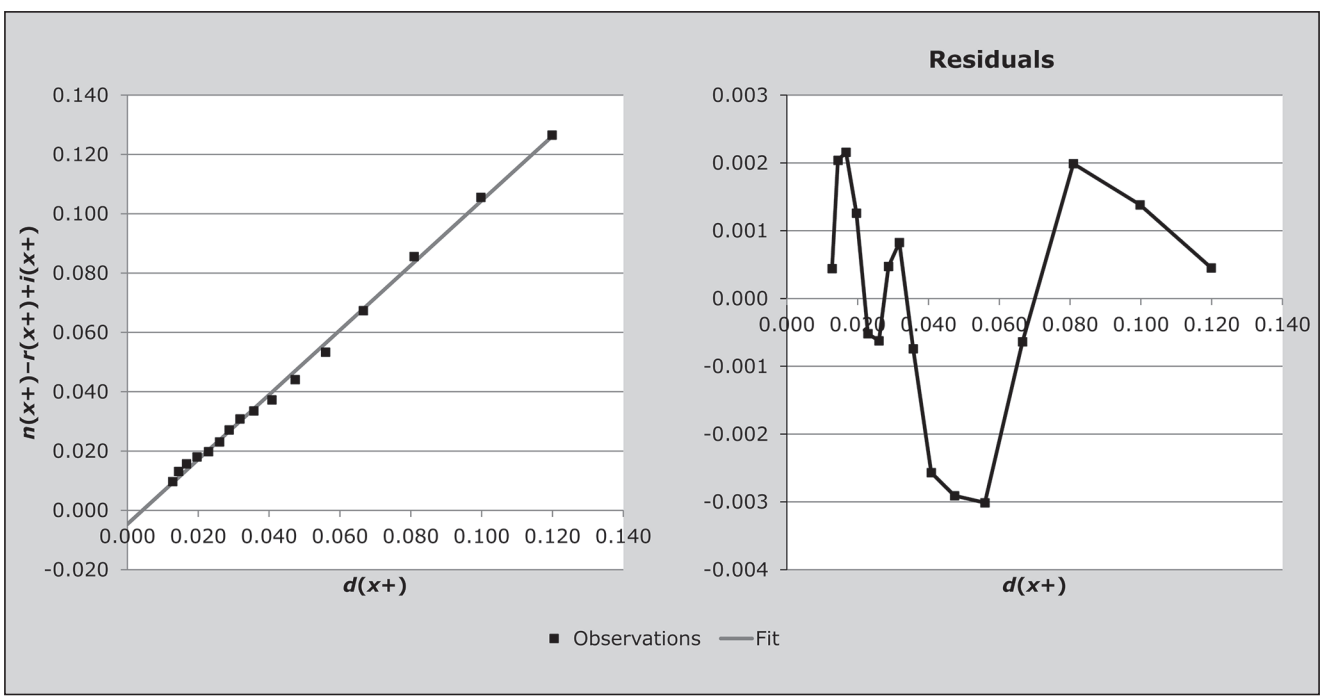
$$d(20+) = \frac{1318246}{67053861} = 0.01966.$$

The partial growth rate less the partial net in-migration rate is shown in column 5 of Table 24.3 and is calculated for age 20, for example, using the cumulated populations given in columns 2 and 3 of Table 24.3 and cumulated net in-migration given in the last column of Table 24.2 as follows:

*TABLE 24.3 Calculation of the cumulated populations, deaths and migrants, South African males, 2001–2007*

| Age | $PYL(x+)$ | $N(x)$ | $b(x+)$ | $r(x+)-i(x+)$ | $d(x+)=X$ | $b(x+)-r(x+)+i(x+)=Y$ | $a+bx$ | Residuals $y-(a+bx)$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 119,775,275 | | | #N/A | 0.00000 | | –0.0047 | |
| 5 | 107,136,837 | 2,554,810 | 0.02385 | 0.01413 | 0.01279 | 0.00972 | 0.0093 | 0.0004 |
| 10 | 93,793,458 | 2,611,355 | 0.02784 | 0.01472 | 0.01445 | 0.01312 | 0.0111 | 0.0020 |
| 15 | 80,461,835 | 2,715,670 | 0.03375 | 0.01805 | 0.01670 | 0.01570 | 0.0135 | 0.0022 |
| 20 | 67,053,861 | 2,577,889 | 0.03845 | 0.02042 | 0.01966 | 0.01803 | 0.0168 | 0.0013 |
| 25 | 55,129,886 | 2,212,329 | 0.04013 | 0.02033 | 0.02291 | 0.01980 | 0.0203 | –0.0005 |
| 30 | 44,604,915 | 2,020,991 | 0.04531 | 0.02223 | 0.02602 | 0.02308 | 0.0237 | –0.0006 |
| 35 | 35,344,071 | 1,682,498 | 0.04760 | 0.02049 | 0.02871 | 0.02712 | 0.0266 | 0.0005 |
| 40 | 27,342,320 | 1,469,826 | 0.05376 | 0.02294 | 0.03178 | 0.03082 | 0.0300 | 0.0008 |
| 45 | 20,537,160 | 1,249,916 | 0.06086 | 0.02735 | 0.03569 | 0.03352 | 0.0343 | –0.0007 |
| 50 | 15,001,678 | 992,684 | 0.06617 | 0.02891 | 0.04080 | 0.03726 | 0.0398 | –0.0026 |
| 55 | 10,574,924 | 790,897 | 0.07479 | 0.03071 | 0.04737 | 0.04408 | 0.0470 | –0.0029 |
| 60 | 7,221,483 | 558,171 | 0.07729 | 0.02396 | 0.05595 | 0.05334 | 0.0564 | –0.0030 |
| 65 | 4,716,866 | 448,343 | 0.09505 | 0.02773 | 0.06659 | 0.06732 | 0.0680 | –0.0006 |
| 70 | 2,857,463 | 290,826 | 0.10178 | 0.01619 | 0.08093 | 0.08559 | 0.0836 | 0.0020 |
| 75 | 1,585,932 | 208,577 | 0.13152 | 0.02599 | 0.09977 | 0.10553 | 0.1041 | 0.0014 |
| 80 | 787,071 | 117,144 | 0.14884 | 0.02230 | 0.11989 | 0.12654 | 0.1261 | 0.0005 |
| 85 | 304,201 | | | | | | | |

*FIGURE 24.1 Diagnostic plots, South African males, 2001–2007*



$$r(20+)-i(20+)=\frac{13276662-11813832-93766}{67053861}$$
$$=0.02042.$$

***Step 6: Plot graph of $b(x+)-r(x+)+i(x+)$ against $d(x+)$, and examine to decide on the range over which the line should be fitted***

In order to plot the graph and fit the line to all of the data points, one starts by setting the lower age to 5 and the upper age to 84 (since the open interval for these data is 85+). The values of $b(x+)-r(x+)+i(x+)$ plotted against $d(x+)$ are shown in Figure 24.1.

Inspection of the diagnostic plots in Figure 24.1 suggests that the points lie fairly close to the fitted straight line, indicating that there is little migration which has not been accounted for. Thus there is little reason to alter the age range over which the line is fitted. Thus, as might be expected, increasing the minimum age has very little effect on the estimate of completeness of 92 per cent. Likewise, even though the results may be affected to some extent by a falling off of completeness at the older ages (see the application of the Synthetic Extinct Generations method to these data), the estimate is little changed by excluding the last or the last two points (i.e. reducing the upper age of age interval used to fit the data). Dropping further points, however, increases completeness to implausible levels which suggests that the data (probably the population data) are far from perfect.

***Step 7: Fit line and estimate completeness, c***

The coefficients of the straight line fitted to the points in Figure 24.1 are estimated as follows:

$$b=\frac{\sigma_y}{\sigma_x}=\frac{0.03483}{0.03193}=1.0907$$
$$a=0.0427-1.0878\times 0.045=-0.00626.$$

The relative completeness of enumeration of the census populations is estimated as follows:

$$\frac{k_1}{k_2}=\exp\left(-0.00467\times 5.3541\right)=0.9753.$$

Thus $k_2>k_1$ and so we assume $k_2 = 1$ and hence $k_1 = 0.9753$ (i.e. the first population is undercounted relative to the second by some 2.5 per cent).

The completeness of reporting of deaths, $c$, is 92 per cent (relative to the 2007 count), calculated as follows:

$$c=\frac{\exp\left(-0.004675\times 5.3541\right)}{1.0907}=0.9175.$$

***Step 8: Estimate mortality rates adjusted for incompleteness of reporting of deaths***

The adjusted population as at the first census date is the enumerated population given in column 2 of Table 24.2 divided by $k_1$. For example the adjusted population for age 20 is

$$\frac{2099417}{0.9753}=2152629.$$

The adjusted population at the second census date is the enumerated population given in column 3 of Table 24.2 divided by $k_2$. Since, by assumption, $k_2 = 1$, these numbers are the same as those given in column 3 of Table 24.2.

Next the deaths are adjusted for incompleteness by dividing the number of reported deaths in each age group shown in column 4 of Table 24.2 by the estimate of completeness. These numbers are shown in column 4 of Table 24.4. For example, for age 20 the number is derived from the number of reported deaths, 54 960, as follows:

$$\frac{54960}{0.9175}=59946.$$

The adjusted person-years of life lived (column 5 of Table 24.4) is the geometric average of the populations in columns 2 and 3 of Table 24.4 multiplied by the length (in years) of the period between the censuses, which in this case is 5.3541 years. For age 20 this is

$$PYL(20,5)=5.3541\left(2152629\times 2362519\right)^{1/2}=12074140.$$

The mortality rates adjusted for incompleteness of reporting of deaths (column 6 of Table 24.4) are derived by dividing the adjusted deaths by the adjusted person-years of life lived. For example, for the 20–24 age group the adjusted rate is calculated as follows:

$$\frac{59946}{12074140}=0.0050\cdot$$

***Step 9: Smooth using relational logit model life table***

Estimates of probabilities of people aged $x$ dying in the next 5 years, ${}_5q_x$, estimated from the adjusted rates of mortality which appear in column 6 of Table 24.4, are shown in the second column of Table 24.5. For example, the probability

*TABLE 24.4 Calculation of adjusted mortality rates, South African males, 2001–2007*

| Age | Adjusted ${}_5N_x(t_1)$ | Adjusted ${}_5N_x(t_2)$ | Adjusted ${}_5D_x$ | Adjusted $PYL(x,5)$ | Adjusted ${}_5m_x$ |
|---|---|---|---|---|---|
| 0 | | | | | |
| 5 | 2,486,532 | 2,560,642 | 16,979 | 13,510,001 | 0.0013 |
| 10 | 2,582,831 | 2,452,339 | 12,224 | 13,474,797 | 0.0009 |
| 15 | 2,515,334 | 2,553,293 | 27,784 | 13,568,508 | 0.0020 |
| 20 | 2,152,629 | 2,362,519 | 59,946 | 12,074,140 | 0.0050 |
| 25 | 1,947,414 | 2,033,165 | 112,129 | 10,653,675 | 0.0105 |
| 30 | 1,635,041 | 1,875,483 | 158,796 | 9,375,725 | 0.0169 |
| 35 | 1,478,197 | 1,548,185 | 159,137 | 8,099,564 | 0.0196 |
| 40 | 1,265,085 | 1,306,900 | 148,269 | 6,884,383 | 0.0215 |
| 45 | 992,273 | 1,104,294 | 131,988 | 5,604,563 | 0.0236 |
| 50 | 789,134 | 888,042 | 121,242 | 4,482,045 | 0.0271 |
| 55 | 566,403 | 708,812 | 105,641 | 3,392,442 | 0.0311 |
| 60 | 455,861 | 491,871 | 98,089 | 2,535,277 | 0.0387 |
| 65 | 312,561 | 394,305 | 90,359 | 1,879,609 | 0.0481 |
| 70 | 238,500 | 241,976 | 79,663 | 1,286,217 | 0.0619 |
| 75 | 139,925 | 163,112 | 69,665 | 808,863 | 0.0861 |
| 80 | 93,159 | 87,698 | 52,533 | 483,940 | 0.1086 |
| 85 | 47,084 | 70,299 | 50,387 | 308,032 | 0.1636 |

*TABLE 24.5 Calculation of smoothed mortality rates using a relational logit model life table, South African males, 2001–2007*

| Age | ${}_5q_x$ | $l_x/l_5$ | Obs. $Y(x)$ | AIDS Cdn. $l^s(x)$ | Cdn. $Y^s(x)$ | Fitted $Y(x)$ | Fitted $l(x)$ | $T(x)$ | Smooth ${}_5m_x$ |
|---|---|---|---|---|---|---|---|---|---|
| 0 | | | | | | | | | |
| 5 | 0.0063 | 1 | | 1.0000 | | | 1 | 50.898 | 0.0033 |
| 10 | 0.0045 | 0.9937 | –2.5333 | 0.9785 | –1.9081 | –2.0551 | 0.9839 | 45.938 | 0.0029 |
| 15 | 0.0102 | 0.9892 | –2.2605 | 0.9632 | –1.6326 | –1.7292 | 0.9695 | 41.055 | 0.0025 |
| 20 | 0.0245 | 0.9792 | –1.9250 | 0.9512 | –1.4853 | –1.5550 | 0.9573 | 36.238 | 0.0043 |
| 25 | 0.0513 | 0.9552 | –1.5293 | 0.9324 | –1.3120 | –1.3500 | 0.9370 | 31.502 | 0.0089 |
| 30 | 0.0812 | 0.9062 | –1.1339 | 0.8969 | –1.0818 | –1.0777 | 0.8962 | 26.919 | 0.0157 |
| 35 | 0.0936 | 0.8326 | –0.8019 | 0.8420 | –0.8365 | –0.7875 | 0.8285 | 22.608 | 0.0204 |
| 40 | 0.1022 | 0.7546 | –0.5616 | 0.7794 | –0.6311 | –0.5446 | 0.7482 | 18.666 | 0.0237 |
| 45 | 0.1112 | 0.6775 | –0.3711 | 0.7148 | –0.4593 | –0.3414 | 0.6644 | 15.134 | 0.0241 |
| 50 | 0.1267 | 0.6021 | –0.2072 | 0.6560 | –0.3228 | –0.1799 | 0.5890 | 12.001 | 0.0230 |
| 55 | 0.1445 | 0.5259 | –0.0518 | 0.6048 | –0.2127 | –0.0497 | 0.5248 | 9.216 | 0.0254 |
| 60 | 0.1764 | 0.4499 | 0.1005 | 0.5530 | –0.1064 | 0.0760 | 0.4621 | 6.749 | 0.0332 |
| 65 | 0.2146 | 0.3705 | 0.2649 | 0.4918 | 0.0163 | 0.2212 | 0.3912 | 4.616 | 0.0497 |
| 70 | 0.2682 | 0.2910 | 0.4452 | 0.4119 | 0.1781 | 0.4125 | 0.3047 | 2.876 | 0.0709 |
| 75 | 0.3543 | 0.2130 | 0.6535 | 0.3178 | 0.3819 | 0.6536 | 0.2130 | 1.582 | 0.0998 |
| 80 | 0.4269 | 0.1375 | 0.9180 | 0.2173 | 0.6408 | 0.9598 | 0.1279 | 0.730 | 0.1459 |
| 85 | #N/A | 0.0788 | 1.2293 | 0.1201 | 0.9959 | 1.3799 | 0.0595 | 0.261 | 0.2070 |

of a 20-year old woman dying before reaching age 25 is calculated as follows:

$$ {}_5q_{20} = \frac{5 \times 0.005}{1 + 2.5 \times 0.005} = 0.02452 . $$

The life table proportions of five-year olds alive at age $x+5$ estimated from the proportion alive at age $x$ using these values appear in column 3 of Table 24.5. For example, the proportion alive at age 25 is calculated as follows:

$$ l_{25} = 0.9792\left(1 - 0.02452\right) = 0.9552 . $$

The logit transformations of the proportions surviving appear in column 4 of Table 24.5. For example, the logit transformation of the $l_{20}$ is calculated as follows:

$$ \gamma_{20} = 0.5 \ln\left(\frac{1 - 0.9792}{0.9792}\right) = -1.925 . $$

The logit transformation of the conditional life table for males based on the AIDS life table with $e_0$=50 in column 5 of Table 24.5 appears in column 6 of Table 24.5. As can be seen from Figure 24.2, the AIDS model does not fit the data particularly well, but fits better than any table which does not reflect the impact of HIV on mortality.

The coefficients, $\alpha$ and $\beta$ are determined as the intercept and slope of the straight line fitted to the logit transformations in columns 4 and 6 of Table 24.5 over the range of ages chosen by the user (45 and 80 in this example), namely 0.2019 and 1.1828 respectively.

These coefficients are then applied to the logit transformation of the conditional model life table to produce the fitted logits in column 7 of Table 24.5. Thus, for example, the fitted logit at age 20 is calculated as follows:

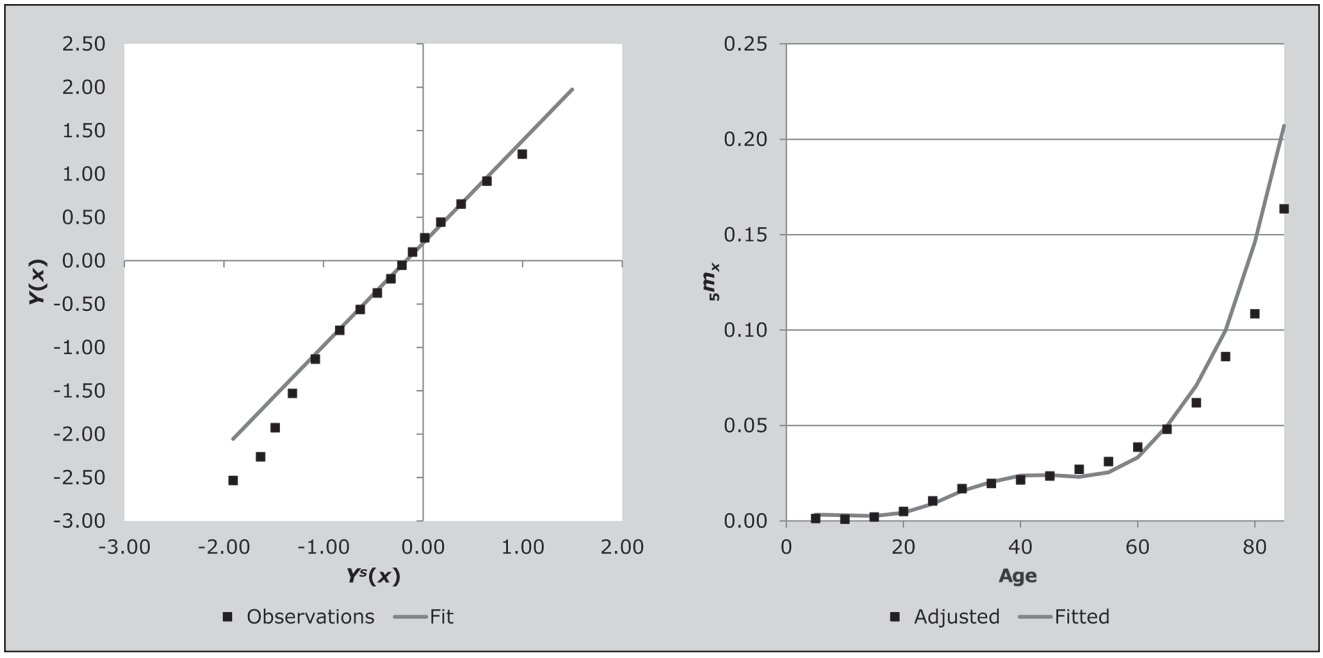
$$ \gamma_{20}^{fitted} = 0.2019 + 1.1828 \times \left(-1.4853\right) = -1.555 . $$

These values are then used to produce the fitted life table in column 8 of Table 24.5. For example, the value at age 20 is calculated as follows:

$$ l_{20}^{fitted} = \frac{1}{1 + \exp\left(2 \times \left(-1.555\right)\right)} = 0.9573 . $$

The conditional years of life lived, $T_x$, which appear in column 9 of Table 24.5 are then calculated from the fitted

***FIGURE 24.2*** *Observed logits and adjusted mortality rates against expected derived from the male AIDS model life table, South African males, 2001–2007*



life table and these numbers are then used to produce the smoothed mortality rates which appear in column 10 of Table 24.5. For example, for age 80

$$T_{80} = 0.261 + \frac{5}{2}(0.1279 + 0.0595) = 0.7299$$

$$_5m_{80}^{fitted} = \frac{0.1279 - 0.0595}{0.7299 - 0.261} = 0.1459 .$$

## DIAGNOSTICS, ANALYSIS AND INTERPRETATION

### Checks and validation

The estimate of completeness is 92 per cent. The first check on this result is a comparison with the results for the opposite sex. For example, applying the same method as described above for men to the data for women during the same period (*GGB_South Africa_females*) gives an estimate of completeness of 89 per cent. Past research (Dorrington, Moultrie and Timæus 2004) leads to the expectation that the estimates should be similar, so the results are sufficiently close to validate the estimates.

A second check on the results is to compare them with the result from the Synthetic Extinct Generations method (*SEG_South Africa_males*), which estimated the completeness of death reporting over the age range 5 to 84 to be 94 per cent, which is also sufficiently close to validate the results.

A third check is to compare estimates of various key indicators of mortality with those from other sources, such as previous estimates for the country or the *World Population Prospects* (UN Population Division 2011). The estimate of $_{45}q_{15}$ from the observed mortality rates after adjusting for incompleteness is 52.3 per cent, while the estimate of $_{45}q_{15}$ from the WPP for the period 2000–2005 is 52.9 per cent, again suggesting little reason to question the results.

As a matter of interest, application of the Brass Growth Balance method to these data (estimating the population in the middle of the period as the average of the two survey populations) provides an estimate of completeness, using the same age range, of 85 per cent. Increasing the minimum age of range of the data used to fit the straight line to 35 increases the estimate to 88 per cent, still somewhat lower than the estimate of 92 per cent produced above.

### Interpretation

As mentioned already, when deciding on the age range over which to fit the straight line, each of the open intervals from 85+ down to 75+ produced virtually the same estimate of the completeness of death reporting. However, below 75+ the estimates increase to 100 per cent for 70+, 105 per cent for 65+ and 108 per cent for 60+. Even though it is probable that the census and survey underestimate the number of men, the undercount is likely to have been concentrated among young adults and is unlikely to have been so great as to raise the completeness of reporting of the deaths relative to the estimate of the population to more than 100 per cent. Moreover, other things being equal, the lower the age of the open interval the less robust the estimate of completeness. Thus, the lower estimates obtained from open-ended age groups for higher ages are preferred.

## METHOD-SPECIFIC ISSUES WITH INTERPRETATION

### Source of reported deaths

Generally there are two sorts of problems with the deaths data: those that lead to under/over coverage that is constant by age, which is precisely what the method is intended to address, and those which lead to differential coverage by age, which can distort the estimates. Although the general approach remains essentially the same irrespective of the source of the death data, different sources of death data are prone to different biases which might impact on the interpretation of the results. These are illustrated by way of particular examples, but, in general terms, the analyst needs to look out for the following biases in the death data.

#### *1) Vital registration*

If the proportionate split of the population between urban and rural (or appropriate proxies) areas differs significantly by age and the completeness of reporting of deaths in urban areas is significantly higher than it is in rural areas, then the assumption that completeness is independent of age is likely to be violated by a falling off of completeness with age at ages over 50 if a proportion of people move from urban to rural areas on retirement. If ignored, this violation is likely to lead to an underestimate of the average level of completeness.

#### *2) Deaths reported by households*

The data are subject to four potential problems:

- If a significant proportion of households dissolve on the death of a key person (e.g. the sole breadwinner), then the deaths of such people go unreported, leading to a violation of the assumption that completeness is invariant with age.

If a significant proportion of deaths in some age groups are of individuals who do not live in private households (for example, they live in homes for the elderly), the breach of the assumption could be even more severe. However, this is not an issue in most developing countries.

- In situations where young adults leave the home they grew up in to work in urban areas, it is possible that they are regarded as being members of more than one household (or of neither household) and their deaths could be reported more than once (or not at all), again leading to a violation of the assumption of constant reporting of deaths by age. In this case one can limit the impact by ignoring the data below a specific age in determining completeness.
- Reference period error: Since there is often confusion about the exact period for which deaths are to be reported, not to mention uncertainty about exact dates of death, it is possible for there to be overall under- or over-reporting of deaths. Provided one can assume that this is independent of the age of the deceased, this distortion will be accounted for in the estimate of completeness and is not a problem for estimating mortality rates.
- The reference period covers a small proportion of the intercensal period. For example, it is common for households to be asked to report on deaths only for the year preceding the census. Not only might such a short period result in significant random fluctuation, but there is a problem that one does not have an estimate of the population at the start of this reference period. How one might deal with this is illustrated in the examples given, but if one has, in addition, deaths reported by households at the first census, one can use the two sets of data on deaths to estimate the number of deaths during the intercensal period, as was discussed above. However, since the question asking households to report on deaths in the previous year was used relatively seldom before the 2010 round of censuses, one may only have the single set of data on deaths. In this case, provided there are no reasons for assuming that the age pattern of mortality has changed rapidly over the period, it is recommended that one calculates the age-specific death rates for the year and applies these to the person-years of life lived for the interval to get an estimate of deaths for the period. If there are reasons for suspecting that mortality has changed rapidly, for example due to HIV/AIDS, then this adaptation is likely to underestimate or overestimate the mortality and the use of death distribution methods is not recommended.

***3) Deaths recorded in health facilities***

Little is known about how well this source of data works. However, it can be expected that completeness would depend on the distribution of health services from which the data have been gathered, and in many developing countries such services are likely to be concentrated in urban areas. So, again, if the proportion of the population living in urban rather than rural areas varies with age, then completeness cannot be assumed to be independent of age. It is also possible that certain causes will predominate in facilities and if these causes are significant, and age-related, this could lead to a further violation of the assumption of constant completeness by age.

## EXAMPLES USING DEATHS REPORTED BY HOUSEHOLDS IN A CENSUS/SURVEY

The examples below use the same data as used in the *GGB_South Africa_males* and *GGB_South Africa_females* workbooks with the exception that instead of using the vital registration as the source of the death data, deaths are estimated from deaths reported by households in the 2001 Census and the 2007 Community survey as having occurred in the year preceding the census/survey. These numbers are given in Table 24.6.

The numbers of deaths occurring between the date of the Census (midnight between 9 and 10 October 2001) and the survey (assumed to be midnight between 14 and 15 February 2007) are estimated using the *Estimating deaths_South Africa_males_hhd* and the *Estimating deaths_South Africa_females_hhd* spreadsheets.

Applying the Generalized Growth Balance method to these data for males (in the *GGB_South Africa_males_hhd* workbook), suggests that these estimates of the number of deaths are more or less as completely reported as the vital registration. However, they estimate ${}_{45}q_{15}$ at 54.3 per cent which is slightly higher than that produced using registered deaths. Applying the Generalized Growth Balance method to these data for females (in the *GGB_South Africa_females_hhd* workbook), suggests that the deaths of women reported by households are far less complete than the registered deaths. It also estimates ${}_{45}q_{15}$ at 50.3 per cent, which is much higher (and less plausible relative to the probability for males) than the 42 per cent produced using registered deaths.

*TABLE 24.6 Deaths reported by households to have occurred in the year preceding census/survey, South Africa*

| | 2001 Census | | 2007 Community Survey | |
|---|---|---|---|---|
| Age | Males | Females | Males | Females |
| 0–4 | 35,873 | 32,096 | 48,322 | 44,418 |
| 5–9 | 3,868 | 3,155 | 4,505 | 5,216 |
| 10–14 | 2,590 | 2,284 | 3,442 | 3,259 |
| 15–19 | 5,628 | 5,122 | 8,246 | 7,878 |
| 20–24 | 10,976 | 13,246 | 16,360 | 21,702 |
| 25–29 | 17,787 | 19,727 | 27,551 | 35,840 |
| 30–34 | 20,038 | 18,292 | 34,832 | 42,576 |
| 35–39 | 19,816 | 15,521 | 38,061 | 34,809 |
| 40–44 | 17,417 | 12,124 | 33,604 | 28,823 |
| 45–49 | 15,840 | 10,105 | 27,829 | 20,973 |
| 50–54 | 15,077 | 9,144 | 28,223 | 18,891 |
| 55–59 | 12,781 | 7,755 | 22,868 | 13,118 |
| 60–64 | 13,428 | 10,367 | 18,775 | 14,912 |
| 65–69 | 11,820 | 10,195 | 17,532 | 14,298 |
| 70–74 | 11,885 | 10,809 | 14,879 | 14,645 |
| 75–79 | 8,794 | 8,393 | 12,966 | 14,151 |
| 80–84 | 7,484 | 9,371 | 9,204 | 12,063 |
| 85+ | 7,115 | 12,389 | 11,735 | 18,178 |

The reason for the much poorer performance of the method when applied to deaths of women reported by households can be seen by a comparison of the estimated numbers of deaths for the period derived from deaths reported by households to the numbers expected after correcting the vital registration for incompleteness of reporting, as shown in Table 24.7. From this we see that there is a significant decline in completeness of reporting of deaths of women by households with age from age 55, probably as the result of the disintegration of households on the death of these women, usually because these households were headed by the women who died.

There is also evidence of over-reporting of deaths below age 30 for males and 25 for females, possibly because their deaths are reported by more than one household.

To simulate the situation where only the most recent census asked about deaths in the previous year, the number of deaths in each age group between the times of the 2001 Census and the 2007 Community Survey using only the deaths reported by households in the 2007 Community Survey are estimated as follows:

$$\frac{{}_5D_x(t_2)}{{}_5N_x(t_2)}(t_2-t_1)\left({}_5N_x(t_1)\times {}_5N_x(t_2)\right)^{1/2}.$$

Applying the method to these estimates of the deaths produces estimates of ${}_{45}q_{15}$ of 58.6 per cent for males and 57.8 per cent for females. Unlike the previous estimates, these are estimates of mortality in the year preceding the second census/survey. They might therefore be expected to be higher than those for the whole period, since mortality has been increasing over the period due to HIV/AIDS. However, as might also be expected, deriving an estimate from a single year of deaths (derived, in addition, in this case from a relatively small sample survey) produces far less reliable estimates, particularly in the case (for these data) of females. Alternative estimates (Bradshaw, Dorrington and Laubscher 2012) suggest that for 2006 the correct probabilities should be closer to 55 per cent for males and 45 per cent for females.

## DETAILED DESCRIPTION OF METHOD

### Mathematical exposition

The General Growth Balance method follows the same logic as Brass's Growth Balance method (Brass 1975), which had its origins in work by Carrier (1958), who first proposed a way of estimating mortality from the age distribution of deaths. The method derives from the simple relationship found in the balancing equation for a population (assumed for convenience of explanation to be) closed to migration. In such a population, the number of people in the population at time $t_2$ = the number at time $t_1$ plus the births that have occurred between time $t_1$ and $t_2$ less the deaths that have occurred between times $t_1$ and $t_2$, i.e. ${}_\infty N_0(t_2) = {}_\infty N_0(t_1) + B - D$, where $B$ and $D$ are the births and deaths, respectively, that occurred between times $t_1$ and $t_2$. This equation can be generalized to hold for any population aged $x$ and older, provided we have an estimate of the number of people who turned $x$ (i.e. joined the age interval through aging) between the times $t_1$ and $t_2$, $N_x$, and the number of deaths aged $x$ and older that occurred between times $t_1$ and $t_2$, ${}_\infty D_x$. Thus

$$ {}_\infty N_x(t_2) = {}_\infty N_x(t_1) + N_x - {}_\infty D_x. \qquad (1)$$

If we rewrite Equation 1 as

$$ {}_\infty N_x(t_2) - {}_\infty N_x(t_1) = N_x - {}_\infty D_x$$

and divide through by the person-years of exposure between times $t_1$ and $t_2$, $\int_{t_1}^{t_2} {}_{\infty}N_x(t)dt$, one can express this balance equation in terms of rates, i.e.

$$r(x+) = b(x+) - d(x+), \qquad (2)$$

where

$$r(x+) = \frac{{}_{\infty}N_x(t_2) - {}_{\infty}N_x(t_1)}{\int_{t_1}^{t_2} {}_{\infty}N_x(t)dt},$$

$$b(x+) = \frac{N_x}{\int_{t_1}^{t_2} {}_{\infty}N_x(t)dt},$$

and

$$d(x+) = \frac{{}_{\infty}D_x}{\int_{t_1}^{t_2} {}_{\infty}N_x(t)dt}.$$

$b(x+)$ and $d(x+)$ are often referred to as *partial or segmental* birth and death rates, respectively.

These relationships only hold if there is complete and accurate recording of birthdays and deaths by age between times $t_1$ and $t_2$, and counting of the population by age at times $t_1$ and $t_2$.

Now, suppose that instead of accurate data only a proportion (the same for all ages) of deaths are reported, and only a (different) proportion (the same for all ages) of each census population, are counted. Suppose further that, instead of the true values ${}_{\infty}N_x(t_1)$, ${}_{\infty}N_x(t_2)$ and ${}_{\infty}D_x$, we have reported values ${}_{\infty}N_x^r(t_1)$, ${}_{\infty}N_x^r(t_2)$ and ${}_{\infty}D_x^r$ such that ${}_{\infty}N_x^r(t_1) = k_1 \cdot {}_{\infty}N_x(t_1)$, ${}_{\infty}N_x^r(t_2) = k_2 \cdot {}_{\infty}N_x(t_2)$ and ${}_{\infty}D_x^r = c \cdot {}_{\infty}D_x$.

Then, if we use the following approximations:

$$r(x+) \approx \frac{\ln\left(\frac{{}_{\infty}N_x(t_2)}{{}_{\infty}N_x(t_1)}\right)}{t_2 - t_1},$$

$$N_x \approx (t_2 - t_1)\frac{\sqrt{{}_nN_{x-n}(t_1) \times {}_nN_x(t_2)}}{n}$$

*TABLE 24.7 Ratio of estimates of deaths derived from deaths reported by households to the expected numbers of deaths, South Africa*

| | Males | | | Females | | |
|---|---|---|---|---|---|---|
| Age | Reported | Expected | Ratio | Reported | Expected | Ratio |
| 0–4 | | | | | | |
| 5–9 | 22,683 | 16,979 | 134% | 22,995 | 14,575 | 158% |
| 10–14 | 16,462 | 12,224 | 135% | 15,173 | 10,349 | 147% |
| 15–19 | 38,013 | 27,784 | 137% | 35,666 | 26,874 | 133% |
| 20–24 | 74,934 | 59,946 | 125% | 95,993 | 84,611 | 113% |
| 25–29 | 124,403 | 112,129 | 111% | 152,718 | 154,437 | 99% |
| 30–34 | 150,792 | 158,796 | 95% | 166,488 | 170,680 | 98% |
| 35–39 | 159,016 | 159,137 | 100% | 137,837 | 141,399 | 97% |
| 40–44 | 140,172 | 148,269 | 95% | 111,910 | 115,746 | 97% |
| 45–49 | 120,016 | 131,988 | 91% | 85,284 | 93,408 | 91% |
| 50–54 | 118,989 | 121,242 | 98% | 76,941 | 81,793 | 94% |
| 55–59 | 97,977 | 105,641 | 93% | 57,353 | 72,131 | 80% |
| 60–64 | 88,088 | 98,089 | 90% | 69,220 | 78,877 | 88% |
| 65–69 | 80,451 | 90,359 | 89% | 67,007 | 86,099 | 78% |
| 70–74 | 72,827 | 79,663 | 91% | 69,536 | 93,404 | 74% |
| 75–79 | 59,632 | 69,665 | 86% | 61,942 | 88,314 | 70% |
| 80-84 | 45,365 | 52,533 | 86% | 58,410 | 77,084 | 76% |
| 85+ | 51,779 | 50,387 | 103% | 83,753 | 108,002 | 78% |

and

$$\int_{t_1}^{t_2} {}_{\infty}N_x(t)\,dt \approx (t_2 - t_1)\left({}_{\infty}N_x(t_1) \times {}_{\infty}N_x(t_2)\right)^{1/2},$$

then

$$r(x+) = r^r(x+) + \frac{\ln\left(\frac{k_1}{k_2}\right)}{t_2 - t_1},$$

where

$$r^r(x+) = \frac{\ln\left(\frac{{}_{\infty}N_x^r(t_2)}{{}_{\infty}N_x^r(t_1)}\right)}{t_2 - t_1},$$

$$b(x+) = b^r(x+) = \frac{(t_2 - t_1)\left({}_nN_{x-5}^r(t_1) \times {}_nN_x^r(t_2)\right)^{1/2} / n}{(t_2 - t_1)\left({}_{\infty}N_x^r(t_1) \times {}_{\infty}N_x^r(t_2)\right)^{1/2}},$$

and

$$d(x+) = \left(\frac{k_1 k_2}{c}\right) d^r(x+),$$

where

$$d^r(x+) = \frac{{}_{\infty}D_x^r}{(t_2 - t_1)\left({}_{\infty}N_x^r(t_1) \times {}_{\infty}N_x^r(t_2)\right)^{1/2}},$$

and Equation 2 becomes

$$r^r(x+) + \frac{\ln\left(\frac{k_1}{k_2}\right)}{t_2 - t_1} = b^r(x+) - d^r(x+)\left(\frac{k_1 k_2}{c}\right),$$

i.e.

$$b^r(x+) - r^r(x+) = a + b d^r(x+)$$

where

$$a = \frac{\ln(k_1/k_2)}{t_2 - t_1} \text{ and } b = \left(\frac{k_1 k_2}{c}\right).$$

From this one can solve for $k_1$, $k_2$ and hence $c$, on the assumption that coverage of the better enumerated census is 100 per cent, by assuming the larger of $k_1$ and $k_2 = 1$.

***Fitting of the straight line***

There are two aspects to determining the straight line that best represents the relationship between the partial birth and death rates, namely, the choice of method and the choice of points used to determine the slope and intercept.

Fitting the straight line using unweighted least squares regression is not recommended since this method gives too much weight to outliers, which tend to be less reliable, particularly at the older ages. Thus it is recommended that one fit the line using a more robust method such as the 'mean' line (i.e. the line defined as that joining the two points represented by the mean of the vertical axis values and the mean of the horizontal axis values of the first half and the second half of the age range) or the 'trimmed mean' line (i.e. the same as the mean line except that the average of the points is a weighted average – weighting the less reliable points, usually at the extremes, less than the other points). These methods are explained in detail in *Manual X* (UN Population Division 1983: 144–145). An alternative is described in more detail in the *UN Manual on Adult Mortality* (UN Population Division 2002: 105–110). It is similar to the 'mean' line, except that one splits the range of points into three equally sized groups,[1] and determines the line that joins the medians of the independent and dependent variables in the lowest third and the highest third of points.

Bhat (2002) points out that each method has its drawbacks and suggests, since it matters not whether the partial birth or partial death rates are treated as dependent variable, that orthogonal regression is the best method for dealing with age misstatement. This reflects both vertical and horizontal distance from the line (by minimizing the orthogonal residual sum of squares (ORSS) = $\sum_i \frac{x_i^2 y_i^2}{x_i^2 + y_i^2}$). Using this method, the $c$, the completeness of the death reporting, is estimated as the ratio of the standard deviation of the partial death rates to the standard deviation of the partial birth rates. The intercept is the mean of the partial birth rates, minus the mean of the partial death rates divided by $c$. This is the approach used in the applications of the Generalized Growth Balance method in the accompanying workbooks.

***Limitations***

This method is less vulnerable to age misreporting than the Synthetic Extinct Generations method. However, the common tendency to exaggerate the age reported at death (relative to that recorded at census) will manifest itself by the plotted points falling off to the right (i.e. below the fitted

1 Where it is not possible to divide the total number of points, $n$, into three equally sized groups then the highest and lowest group are taken as the top and bottom, respectively, int($n$/3)+1 points.

line) over the range of exaggerated ages. This can be catered for by reducing the age of the open interval to the point which removes this effect.

Migration which is not allowed for in the model is likely to affect the young adult population (mainly between 20 and 35) but to have much less effect on deaths, which largely occur in old age. Unaccounted-for immigration will tend to lower the slope and hence lead to an over-estimate of the extent of death registration and an under-estimate of mortality rates. Unaccounted-for emigration will have the opposite effect.

Often one lacks reliable estimates of the net number of migrants by age over the intercensal period. In such situations one could proceed as follows. If the migration is significant and unknown and the points above age 30 lie close to a straight line, one might estimate completeness by fitting the straight line to the data from age 35 and above. If migration is slight, some demographers advocate fitting the straight line to data down to age 5 to limit this distortion, on the assumption that any differences in completeness of reporting of deaths at these younger ages from that of the older ages is unlikely to lead to any major distortions since the mortality is very light between ages 5 and 14. Others (Hill, You and Choi 2009) suggest that provided the migration is not too significant, an improved estimate might be provided by averaging the estimate of completeness produced with that produced by applying the Synthetic Extinct Generations method to the same data. Although using these adaptations probably produces better estimates than simply ignoring migration, there is, unfortunately, little research into the accuracy of the estimated completeness produced by these adaptations.

Fluctuations in the completeness of death registration with age are likely to introduce curvature in the pattern of points. Consequently, one of the strengths of this method is that if the points for successive age boundaries fall on a reasonably straight line, then it is probably reasonable to assume that completeness is constant with respect to age. However, where some but not all the points lie on a straight line, one way of deciding which points to discard is to calculate the segmental growth rate for each successive open interval and then use those points for which the values of $r_{a+}$ are reasonably consistent.

Perhaps the most important limitation of the method is that the plot of partial birth rates against partial death rates is, with the exceptions mentioned above, diagnostically quite limited.

***Extensions***

If the ages were recorded accurately and the assumption of constant census coverage by age held, then the method could be adapted to deal with the situation where completeness of reporting of the deaths was constant only for a limited age range ($x$ to $x+n$)[2] by limiting the age range of the balance equation. Thus Equation 2 would become

$$ {}_nr_x = b(x+) - {}_nd_x - b\big((x+n)+\big), $$

where

$$ {}_nr_x = \frac{{}_nN_x(t_2) - {}_nN_x(t_1)}{\int_{t_1}^{t_2} {}_nN_x(t)dt}, $$

$$ b(x+) = \frac{N_x}{\int_{t_1}^{t_2} {}_\infty N_x(t)dt} $$

and

$$ {}_nd_x = \frac{{}_nD_x}{\int_{t_1}^{t_2} {}_nN_x(t)dt}. $$

The LHS of the analogous regression equation based on observations becomes $b^r(x+) - b^r(x+n+) - {}_nr_x^r$.

Perhaps because data in developing countries are rarely accurate enough, little experience exists with how well this alternative approach works in practice.

## FURTHER READING AND REFERENCES

Analysis of the sensitivity of the method to common data errors and violation of the assumptions is fairly limited. However, the reader is referred to Hill, You and Choi (2009) for an analysis of the assumptions underlying the death distribution methods in the absence of HIV and to Dorrington and Timæus (2008) for an analysis in a population experiencing significant HIV. Murray, Rajaratnam, Marcus *et al.* (2010), in contrast, used stochastic simulations to assess these methods, concluding that the methods were not particularly reliable. However, to date their work has had very limited impact on the use

2 Such as vital registration where completeness might fall off above retirement age, if people retired from urban to rural areas, or deaths reported by households where household might disintegrate on the death of the last adult.

of these methods, possibly because their description of their simulations is short on detail and because their assessment is based on perhaps unrealistically high migration.


Bhat M. 2002. "General Growth Balance method: A reformulation for populations open to migration", *Population Studies* **56**(1):23–34. doi: 10.1080/00324720213798

Blacker J. 1988. *An Evaluation of the Pakistan Demographic Survey*. Karachi: Pakistan Federal Bureau of Statistics.

Bradshaw D, RE Dorrington and R Laubscher. 2012. *Rapid Mortality Surveillance Report 2012*. Cape Town: South African Medical Research Council. http://www.mrc.ac.za/bod/RapidMortality2011.pdf

Brass W. 1975. *Methods for Estimating Fertility and Mortality from Limited and Defective Data.* Chapel Hill NC: Carolina Population Centre.

Carrier NH. 1958. "A note on the estimation of mortality and other population characteristics, given death by age", *Population Studies* **12**(2):149–163. doi: 10.2307/2172187

Dorrington RE, TA Moultrie and IM Timæus. 2004. *Estimation of Mortality using the South African 2001 Census Data*. Monograph 11. Centre for Actuarial Research, University of Cape Town. http://www.commerce.uct.ac.za/care/Monographs/Monographs/Mono11.pdf

Dorrington RE and IM Timæus. 2008. "Death Distribution Methods for Estimating Adult Mortality: Sensitivity Analysis with Simulated Data Errors, Revisited," Paper presented at Population Association of America 2008 Annual Meeting. New Orleans, Louisiana, 17–19 April.

Hill K. 1987. "Estimating census and death registration completeness", *Asian and Pacific Census Forum* 1(3):8–13, 23–24. http://hdl.handle.net/10125/3602.

Hill K, D You and Y Choi. 2009. "Death distribution methods for estimating adult mortality: Sensitivity analysis with simulated data error", *Demographic Research* **21**(Article 9):235–254. doi: 10.4054/DemRes.2009.21.9

Murray CJL, JK Rajaratnam, J Marcus, T Laakso and AD Lopez. 2010. "What can we conclude from death registration? Improved methods for evaluating completeness", *PLoS Med* **7**(4):e1000262. doi: 10.1371/journal.pmed.1000262

Timæus IM. 2004. "Impact of HIV on mortality in Southern Africa: Evidence from demographic surveillance," Paper presented at Seminar of the IUSSP Committee "Emerging Health Threats" HIV, Resurgent Infections and Population Change in Africa. Ougadougou, 12–14 February.

UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

UN Population Division. 2002. *Methods for Estimating Adult Mortality*. New York: United Nations, Department of Economic and Social Affairs, ESA/P/WP.175. http://www.un.org/esa/population/techcoop/DemEst/methods_adultmort/methods_adultmort.html

UN Population Division. 2011. *World Population Prospects: The 2010 Revision, Volume I: Comprehensive Tables*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/313. http://esa.un.org/unpd/wpp/Documentation/pdf/WPP2010_Volume-I_Comprehensive-Tables.pdf

# Chapter 25 The Synthetic Extinct Generations method

*Rob Dorrington*

## DESCRIPTION OF METHOD

Bennett and Horiuchi (1981, 1984) generalized the Preston and Coale method for estimating the completeness of the reporting of deaths relative to an estimate of the population, into what has eventually become known as the Synthetic Extinct Generations (SEG) method. Basically these methods make use of the observation that, in a closed population, the number of people of a given age, $a$, alive at a point in time must equal the number of people who will die from that age forward. The key insight on which SEG methods are based is that the future stream of deaths of a cohort can be replaced by current deaths at each age above $a$ if the effects of future age-specific population growth can be taken into account. The simplest example is a stationary (life table) population, in which future deaths above age $a$ will be equal to current deaths above that age. Somewhat more complex is the case where the population is stable (i.e. a population with an unchanging adult age distribution growing at a constant rate, $r$, each year) and closed to migration. In this case, if the reported data are accurate, the number of deaths at age $x$, $t$ years in the future, will equal the number of deaths at age $x$ currently, multiplied by $e^{rt}$. This is the model underlying the Preston and Coale method.

In the more general case where the population is not stable, an equivalent relationship exists if one replaces $e^{rt}$ by $e^{\int_a^x r(y,t)dy}$ where $r(y,t)$ represent the growth rate of the population aged $y$ at time $t$.

If the deaths reported at time $t$ can be assumed to be reported to the same extent, $c$, at every adult age, then the estimate of the future number of cohort deaths will be underestimated to the same extent. Thus, it is possible to estimate the completeness of reporting of deaths by dividing the sum of the estimates of future cohort deaths derived from the number of deaths at any date by the population at the same date. Mortality rates can then be estimated by dividing the numbers of deaths reported in each adult age group by $c$ and then dividing these numbers by an estimate of the population exposed to risk based on the population used to estimate the partial birth and death rates.

## DATA REQUIREMENTS AND ASSUMPTIONS

### Tabulations of data required

- Number of women (men), by five-year age group, and for open age interval $A$+ (with $A$ as high as possible), at two points in time, typically from the results of two censuses. (See the caveat below concerning the use of surveys rather than censuses.)
- Number of deaths of women (men), by five-year age group, and for open age interval $A$+, over the period between the two censuses or surveys.

### Important assumptions

- The coverage of each census is the same for all ages.
- The completeness of reporting of deaths is the same for all ages above some minimum age (usually age 15).
- The population is closed to migration. Although the method can be adapted to allow for migration, accurate enough estimates of the net numbers of migrants to do so seldom exist. For national populations, net migration is often low enough to ignore, but for situations where migration is significant one needs to take this into account when interpreting results and deciding on an estimate of completeness.

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Before applying this method, you should examine the quality of the data in at least the following dimensions:

- age structure of the population;
- sex structure of the population;
- age structure of the deaths; and

- sex structure of the deaths.

If the reported deaths are for a period other than that between the censuses the numbers that would have been reported in the intercensal period need to be estimated. If one has annual vital registration data, this adjustment involves apportioning deaths in the first and last year of the period. If one has deaths reported by households the year before the dates of each of the first and second censuses, one has to estimate the numbers of deaths by interpolating between these estimates for the intercensal period (using the *Estimating deaths* spreadsheet).

The SEG methodology uses age-specific population growth rates in its calculations. If the completeness of census enumeration varies from one census to the next by a proportionately constant amount at all ages, such growth rates will be biased by a fixed amount, delta. The Generalized Growth Balance methodology explicitly estimates this bias. The SEG methodology does not explicitly estimate delta, but non-zero values of delta result in a linear trend in estimates of completeness of death recording with age. Thus delta can be estimated iteratively by finding the value that produces coverage estimates that are constant by age between selected lower and upper age limits.

## CAVEATS AND WARNINGS

In applying this method, analysts must take particular care with the following.

- The interpretation and estimating processes need to take into account the source of death data (vital registration, reported by households in censuses, or recorded in hospitals) as explained below. Biases associated with the source of death data tend to have more impact on the estimate of completeness from the Synthetic Extinct Generations method than on the Generalized Growth Balance method.
- If applying the method to sub-national geographic areas, the issue of migration typically becomes a greater concern.
- Deciding the age range which is to be used to determine delta (the estimate of the coverage of one census relative to the other). Issues here are whether the best estimate of delta is the intercept determined as a result of applying the Generalized Growth Balance method to the same data (which would be the case, for example, if completeness of reporting of deaths was thought to decrease after retirement, for reasons explained below), and whether to exclude ages below 30 or 35 because of the impact of migration which has not been allowed for specifically.
- Deciding on the age range to use for determining the estimate of completeness. Typically this range might exclude young adults if there is significant unaccounted-for migration, or the elderly if the results suggest that a lower proportion of their deaths are reported than of deaths of younger adults or if age misreporting of the elderly appears to affect ages of the living and the dead differently.
- Ensuring that the solver routine in Excel has run satisfactorily (i.e. has produced a sensible result). Occasionally *Solver* offers a solution which is manifestly too low. In such situations it is best to adjust delta manually in the right direction and apply *Solver* to this new starting value.
- Ensuring that the estimate of life expectancy at the age of the open interval is reasonable. Often data on older people are scanty and particularly prone to errors. Thus estimates of life expectancy based on these data can be implausible (usually over-estimating life expectancy). The higher the age of the open interval, the lower the impact of any error.
- If completeness appears to be less than 60 per cent, then the uncertainty is large and this should be taken into account when interpreting the results.
- It is tempting in a situation in which census data on the age distribution of the population and household deaths are available for only one census, to combine this with data on the age distribution of the population from a sample survey at some earlier or later date. However, for reasons that have not been adequately researched, such a combination of data sources rarely gives satisfactory results.

## APPLICATION OF METHOD

Although technically one could apply this method to data in single year age categories, the data one typically works with are subject to age misstatement, so in practice one usually works with data grouped into five-year age groups. For convenience, since most data are published in this format, the spreadsheet is set up to work with data in the standard five-year groupings. However, as Blacker (1988) has shown, if this grouping fails to remove the effect of digit preference, the method should be adapted to work with an alternative five-year grouping of ages centred on, rather than starting with, ages at which heaping occurs.

***Step 1: If not readily available, estimate the number of deaths reported in the period between the dates of the two estimates of the population***

In the case where one has annual vital registration data, this adjustment involves apportioning deaths in the first and last year of the period to the parts of the year before and after the mean dates of fieldwork of the two inquiries. Unless the age pattern of deaths is changing very rapidly, this will have no effect on the results.

If one lacks data on the number of deaths between the two inquiries but this interval falls between two periods for which one does have such estimates (for example, because each inquiry included a question about deaths in the household during the previous year), one can make use of the *Estimating deaths* spreadsheet. This spreadsheet estimates the number of deaths between two points in time given estimates of deaths over two other periods. To use this spreadsheet, you need the number of deaths divided into five-year age groups for two periods (periods 1 and 2), the start and end dates for each of these periods, and the start date and end date of the period for which one wishes to estimate the number of deaths.

***Step 2: Estimate the growth rates adjusting for migration and differential census coverage***

Age-specific growth rates adjusting for migration and differential census coverage are estimated from the two census populations and the numbers of migrants over the intercensal period by age group as follows:

$$ {}_5r_x = \frac{\ln\left( {}_5N_x(t_2) / {}_5N_x(t_1)\right)}{t_2 - t_1} - \frac{{}_5NM_x}{\left(t_2 - t_1\right)\left( {}_5N_x(t_1) \times {}_5N_x(t_2)\right)^{\frac{1}{2}}} + \delta \cdot $$

where ${}_5N_x(t)$ is the population aged between $x$ and $x+5$ at time $t$, ${}_5NM_x$ is the net number of migrants (in- less out-migrants) aged between $x$ and $x+5$, and $t_1$ and $t_2$ are the times of the two censuses. Delta is the correction for the completeness of one census relative to the other. It is either set equal to the estimate from the Generalized Growth Balance method, or solved for iteratively as explained below.

***Step 3: Estimate the life expectancy at age A and at each five-year age intervals down to 65***

This can be done in one of several ways.

1) Use estimates from an independent source, if reliable estimates are available. Possible sources would be estimates produced by previous research or from population projections such as the World Population Prospects (UN Population Division 2011).
2) Use the estimates derived from the data after applying the Generalized Growth Balance method. The workbook implementing that method produces such estimates as part of the output.
3) Use the ratio of the reported deaths in the age group 10 to 39 last birthday to those in the age group 40 to 59 last birthday (${}_{30}D_{10}/{}_{20}D_{40}$) to determine (by comparison) a level of the West model life table, from which estimates of life expectancy can be read. These estimates are included as part of the workbook implementing this method. Unfortunately, since the West model life table does not reflect mortality resulting from HIV/AIDS, this approach is unsuitable for countries that have significant numbers of AIDS deaths.
4) Solve for the life expectancy iteratively by starting with a reasonable guess such as those estimated from the West table (although in some cases this may not work in countries with significant numbers of AIDS deaths) or from an independent source. Then estimate completeness (as described below), copy the life expectancies based on this level of completeness from the *Life expectancies* spreadsheet of the associated workbook, paste the values of these into the *Method* spreadsheet of the associated workbook and re-estimate completeness. Repeat if necessary until the change to life expectancies is no longer significant. Unfortunately, if there are reasons for suspecting that, even after correcting the rates for incompleteness, mortality is underestimated at the older ages (for example, if there is significant age exaggeration, or relatively higher incompleteness at the older ages) then this approach will overestimate the life expectancies and hence overestimate the overall level of completeness of reporting.

***Step 4: Estimate the number of people who turned x, and the number aged x to x+4 last birthday, from the reported deaths***

The number of people who turned $x$ during the period over which the deaths were reported is estimated from the reported deaths as follows:

$$ \hat{N}_x = \hat{N}_{x+5} \exp\left(5\, {}_5r_x\right) + {}_5D_x \exp\left(2.5\, {}_5r_x\right) $$

and

$$\hat{N}_A = {}_{\infty}D_A\left(\exp\left({}_{\infty}r_A \times e_A\right) - \left({}_{\infty}r_A \times e_A\right)^2 / 6\right)$$

where $A$ is the age at the start of the open interval, ${}_nr_x$ is the annual population growth rate in the age group $x$ to $x+n$ last birthday, and $e_A$ is the life expectancy at age $A$.

The number of people aged $x$ to $x+4$ last birthday during the period over which the deaths were reported is estimated from the numbers who turned $x$ in five-year steps as follows:

$${}_5\hat{N}_x = 2.5\left(\hat{N}_x + \hat{N}_{x+5}\right).$$

***Step 5: Estimate the number of people aged x to x+4 last birthday during the period between the two censuses, from the census populations***

The number of people aged $x$ to $x+4$ during the period over which the deaths are reported is estimated from the census populations by multiplying the geometric mean of the numbers in that age group in the two census populations by the length of the period between the two censuses (measured in years) as follows:

$${}_5N_x = \left(t_2 - t_1\right)\left({}_5N_x(t_1) \times {}_5N_x(t_2)\right)^{1/2}.$$

***Step 6: Calculate the ratios of the estimates derived from deaths to those derived from the census populations***

Two sets of ratios of the estimates derived from the deaths to those derived from the census population are calculated. The first is the ratios in quinquennial age groups, which are calculated directly. The second is the ratios of the numbers from age $x$ to that age of the open interval, $A$, with the numbers of people who turned $x$ to $A-1$ during the period being calculated as the aggregate of the numbers in five-year age groups between ages $x$ and $A-5$. In other words,

$${}_{A-x}\hat{N}_x = \sum_{a=x,5}^{A-5} {}_5\hat{N}_x .$$

***Step 7: Estimate the completeness of reporting of deaths***

In order to determine the level of completeness of reporting one first needs to decide if the growth rates need to be corrected for relative completeness of the population censuses. The interpretation of the plots of the ratios is discussed in more detail below. However, essentially the amount by which the growth rates need to be corrected (delta) for relative completeness of the census populations is identified as the amount which produces the most level set of ratios by age. The *Method* spreadsheet is set up so that *Solver* (Data, Solver, Solve) will find the value of delta that minimizes the absolute deviation from the mean of the ratios over the age range specified by the user.

It is suggested that the intercept, $a$, from the application of the Generalized Growth Balance method to the same data be used as an initial estimate of delta. If this estimate of delta produces a level series of ratios across adult ages but with significant curvature downward at the older ages, this could indicate a fall off of completeness at the older ages. This might be the case if, for example, people retired from urban areas to rural areas, where completeness of registration was lower, or there was a drop off in reporting of deaths by households due to disintegration of some households on the death of a member. In such a situation it is important **not** to set delta to produce a level set of ratios, but rather to use the initial value.

If one is solving for both delta and the life expectancies iteratively, the values of life expectancies will need to be pasted from the *Life expectancies* spreadsheet into the *Method* spreadsheet and a new estimate of delta set. This process may need to be repeated two or three times, until there is no change in the life expectancies.

Finally, one decides on the age range of ratios to be used to determine the completeness. If there is a significant curvature upward at the older ages, this probably indicates age exaggeration, particularly for deaths, and one needs to try and identify an age for the open interval below which the age exaggeration is not significant. If completeness drops off at ages below 35, this could indicate unaccounted for out-migration. If this is suspected then one should exclude these ages from determining delta or completeness.

Completeness is estimated from the quinquennial age group-specific ratios. In order to produce a robust estimate it is calculated as the sum of 50 per cent of the median plus 25 per cent of each of the 75th and 25th percentile of these ratios.

***Step 8: Estimate mortality rates adjusted for incompleteness of reporting of deaths***

In order to compute mortality rates one needs first to correct the census population for relative under enumeration. This is achieved by multiplying the numbers from the first census by

$\exp\left(-\delta\left(t_2 - t_1\right)\right)$, and the numbers from the second census by 1, if delta is less than zero, and the reverse if delta >0.

The adjusted person-years of exposure, $PYL^a(x,5)$, are estimated by multiplying the geometric average of the corrected populations by the length of the period between the censuses (in years) as follows:

$$PYL^a(x,5) = \left(t_2 - t_1\right)\left({}_5N_x(t_1)e^{-\delta(t_2-t_1)} \times {}_5N_x(t_2)\right)^{1/2}, \quad \delta < 0$$

$$PYL^a(x,5) = \left(t_2 - t_1\right)\left({}_5N_x(t_1) \times {}_5N_x(t_2)e^{-\delta(t_2-t_1)}\right)^{1/2}, \quad \delta > 0\,.$$

Next, one needs to adjust the number of deaths for incompleteness by dividing the reported number of deaths by the estimate of completeness, *c*, and dividing this by $PYL^a(x,5)$ to produce mortality rates adjusted for the incompleteness of the reporting of deaths as follows:

$${}_5m_x = \frac{{}_5D_x/c}{PYL^a(x,5)}.$$

If it is decided that completeness fell after retirement then the estimates of the rates at these ages might be improved by replacing *c* by the age group-specific estimates of completeness at these ages. There is an option in the spreadsheet to implement this.

Note that technically one could drop the adjustment for under-coverage of one census relative to the other and still get the same estimates of the mortality rates since the same adjustment is made to both the numerator and the denominator. However, in that case the estimate of completeness is relative to the average of the census populations ignoring the fact that one is undercounted relative to the other.

***Step 9: Smooth using relational logit model life table***

Because the age-specific rates can be quite erratic they need to be graduated (smoothed). This can be achieved by fitting a Brass relational logit function to a sex-specific standard life table which is considered to have the same shape as that generated by the mortality rates of the population being investigated.

The accompanying workbooks contain a spreadsheet (see website) that allows one to produce a smooth set of mortality rates by using a relational logit model fitted to the life table generated by the adjusted mortality rates. The user can choose between the standard from the General family of United Nations model life tables or one from any of the four families of Princeton model life tables. The logit transforms of these tables together with a model life table of a population experiencing an AIDS epidemic (Timæus 2004) appear in the *Models* spreadsheet. This spreadsheet also allows the user to input logit transforms of an alternative life table if there is reason to assume that it has a similar pattern of adult mortality to that of the population being studied.

In order to fit the model, probabilities of people aged *x* dying in the next 5 years, ${}_5q_x$, are estimated from the adjusted rates of mortality as follows:

$${}_5q_x = \frac{5\,{}_5m_x}{1 + 2.5\,{}_5m_x}.$$

From this the life table with a radix of $l_5 = 1$ is calculated as follows:

$$l_{x+5} = l_x\left(1 - {}_5q_x\right).$$

The coefficients, $\alpha$ and $\beta$ are determined by fitting the relational logit model as follows:

$$\gamma_x = \alpha + \beta\gamma_x^s$$

where

$$\gamma_x = 0.5\ln\left(\frac{1 - l_x}{l_x}\right)$$

and superscript '*s*' designates values based on a standard life table.

The fitted life table is then generated from the standard life table using the coefficients $\alpha$ and $\beta$ as follows:

$$\gamma_x^{fitted} = \alpha + \beta\gamma_x^s$$

and

$$l_x^{fitted} = \frac{1}{\exp\left(1 + 2\gamma_x^{fitted}\right)}.$$

The smoothed mortality rates are derived from this life table as follows:

$${}_5m_x^{fitted} = \frac{l_x^{fitted} - l_{x+5}^{fitted}}{T_x - T_{x+5}}$$

and

$${}_\infty m_x^{fitted} = \frac{l_x^{fitted}}{T_x}$$

where

$$T_x = \sum_{x=x,5}^{\omega} \frac{5}{2}\left(l_x^{fitted} + l_{x+5}^{fitted}\right)$$

i.e.

$$T_x = T_{x+5} + \frac{5}{2}\left(l_x^{fitted} + l_{x+5}^{fitted}\right)$$

and $\omega$ is the age above which the life table has no more survivors.

The life expectancies, which are of particular interest if one wants to estimate the life expectancies at the older ages iteratively, are derived as follows:

$$e_x = \frac{T_x}{l_x}.$$

## WORKED EXAMPLE

This example uses data on the numbers of males in the population from the South African Census in 2001 and the Community Survey in 2007, on number of deaths from vital registration for the years 2001 to 2007, and on the net number of migrants estimated from the change in foreign-born counted in the two surveys, less an estimate of the number of South Africans who emigrated between the two surveys. The example appears in the *SEG_South Africa_males* workbook.

### *Step 1: If not readily available, estimate the number of deaths reported in the period between the dates of the two estimates of the population*

The registered deaths for the years 2001 to 2007 for South African males are given in Table 25.1.

The reference time for the Census in 2001 was midnight between 9 and 10 October 2001. The Community Survey took place over a number of weeks in February so we can assume a reference time of midnight between 14 and 15 February 2007. Thus if we assume deaths occur uniformly over the respective calendar years we can apportion the deaths in 2001 and 2007 and add these to the total for the years 2002 to 2006 to get the total number of deaths between the two estimates of the population. For example, for the age group 20–24 the number is calculated as follows:

$$\frac{(22+30+31)}{365}8931 + 51588 + \frac{(31+14)}{365}10875 = 54960.$$

*TABLE 25.1 Calculation of deaths between census dates, South African males, 2001–2007*

| Age | 2001 | 2002–2006 | 2007 | Total between censuses |
|---|---|---|---|---|
| 0–4 | 29,005 | 186,346 | 40,314 | 197,912 |
| 5–9 | 2,118 | 14,733 | 2,854 | 15,566 |
| 10–14 | 1,745 | 10,535 | 2,233 | 11,207 |
| 15–19 | 4,470 | 23,857 | 4,860 | 25,473 |
| 20–24 | 8,931 | 51,588 | 10,875 | 54,960 |
| 25–29 | 16,834 | 96,705 | 18,405 | 102,802 |
| 30–34 | 20,892 | 137,355 | 28,245 | 145,588 |
| 35–39 | 21,068 | 137,502 | 29,258 | 145,900 |
| 40–44 | 19,322 | 128,217 | 26,973 | 135,936 |
| 45–49 | 17,881 | 113,891 | 24,761 | 121,010 |
| 50–54 | 16,883 | 104,508 | 22,790 | 111,157 |
| 55–59 | 14,544 | 90,919 | 21,317 | 96,854 |
| 60–64 | 15,097 | 84,351 | 17,410 | 89,930 |
| 65–69 | 13,011 | 77,680 | 17,878 | 82,843 |
| 70–74 | 14,035 | 68,147 | 13,771 | 73,036 |
| 75–79 | 10,846 | 59,859 | 12,534 | 63,871 |
| 80–84 | 9,161 | 44,986 | 8,872 | 48,163 |
| 85+ | 7,602 | 43,233 | 10,009 | 46,196 |

### *Step 2: Estimate the growth rates adjusting for migration and differential census coverage*

Age-specific growth rates less the net in-migration rate and adjusted for differential census coverage appear in column 6 of Table 25.2. They are calculated for the 20 to 25 age group, for example, using the populations given in columns 2 and 3, the net in-migration given in column 5 of Table 25.2 and delta (estimated below) as follows:

$${}_5r_{20} = \frac{\ln\left(2362519/2099417\right)}{5.3541} - \frac{14803}{5.3541\left(2099417 \times 2362519\right)^{\frac{1}{2}}} + \left(-0.00467\right) = 0.0161$$

where 5.3541 is the time between the census and survey calculated using the YEARFRAC functions in Excel.

### *Step 3: Estimate the life expectancy at age A and five-year age intervals down to 65*

The estimates derived after applying the Generalized Growth Balance method to the same data are as shown in column 2 of Table 25.3.

*TABLE 25.2 Growth rates and estimate of the numbers who turned x and the numbers aged between x and x + 5 derived from the numbers of deaths, South African males, 2001–2007*

| Age | ${}_5N_x(t_1)$ | ${}_5N_x(t_2)$ | ${}_5D_x$ | ${}_5NM_x$ | ${}_5r_x$ | *Est* $N_x$ | *Est* ${}_5N_x$ |
|---|---|---|---|---|---|---|---|
| 0 | 2,223,006 | 2,505,744 | 197,912 | 10,605 | 0.0168 | | |
| 5 | 2,425,066 | 2,560,642 | 15,566 | 2,848 | 0.0053 | 2,304,653 | 11,334,968 |
| 10 | 2,518,985 | 2,452,339 | 11,207 | 5,153 | –0.0101 | 2,229,335 | 11,405,753 |
| 15 | 2,453,156 | 2,553,293 | 25,473 | 16,574 | 0.0016 | 2,332,967 | 11,556,063 |
| 20 | 2,099,417 | 2,362,519 | 54,960 | 14,803 | 0.0161 | 2,289,459 | 10,871,687 |
| 25 | 1,899,275 | 2,033,165 | 102,802 | 4,714 | 0.0076 | 2,059,216 | 9,851,950 |
| 30 | 1,594,624 | 1,875,483 | 145,588 | 13,331 | 0.0242 | 1,881,564 | 8,529,425 |
| 35 | 1,441,657 | 1,548,185 | 145,900 | 9,693 | 0.0074 | 1,530,206 | 7,153,512 |
| 40 | 1,233,813 | 1,306,900 | 135,936 | 7,464 | 0.0050 | 1,331,199 | 6,238,580 |
| 45 | 967,744 | 1,104,294 | 121,010 | 8,719 | 0.0184 | 1,164,233 | 5,276,384 |
| 50 | 769,627 | 888,042 | 111,157 | 9,413 | 0.0199 | 946,320 | 4,242,847 |
| 55 | 552,402 | 708,812 | 96,854 | 4,640 | 0.0405 | 750,818 | 3,191,145 |
| 60 | 444,592 | 491,871 | 89,930 | 5,081 | 0.0122 | 525,640 | 2,332,526 |
| 65 | 304,835 | 394,305 | 82,843 | 4,922 | 0.0407 | 407,371 | 1,662,114 |
| 70 | 232,604 | 241,976 | 73,036 | 4,334 | –0.0007 | 257,475 | 1,106,744 |
| 75 | 136,466 | 163,112 | 63,871 | 2,980 | 0.0249 | 185,223 | 721,856 |
| 80 | 90,856 | 87,698 | 48,163 | 1,662 | –0.0148 | 103,519 | 412,486 |
| 85 | 45,920 | 70,299 | 46,196 | 2,009 | 0.0683 | 61,475 | |

The ratio of the reported deaths in the age group 10 to 39 last birthday to those in the age group 40 to 59 last birthday from column 4 of Table 25.2 is

$$\frac{485930}{464957} = 1.0451 .$$

The life expectancies of the male Princeton West model life table which corresponds to this are determined (from the table in the *Life expectancies* spreadsheet of the workbook) by interpolation and are shown in column 3 of Table 25.3 For example for age 65:

$$e_{65} = 9.65 + \frac{1.045 - 1.012}{1.062 - 1.012}(9.35 - 9.65) = 9.452 .$$

Solving for the life expectancy and delta iteratively by starting with the estimates from the West table produces an estimate of delta (as explained in more detail below) of –0.0066 and the final estimates of life expectancy which appear in column 4 of Table 25.3.

Since the prevalence of HIV/AIDS was high in South Africa one cannot use the estimates derived from the West life tables given in the *Life expectancies* spreadsheet of the workbook in estimating the completeness of reporting of deaths. In addition, since – discussed below – it appears that completeness could be falling with age for ages above age 55, the iterative estimates may not be ideal. Thus for this example, delta is set equal to the intercept, *a*, from the application of the Generalized Growth Balance method to the same data, and the life expectancies are as appear in column 5 of Table 25.3.

*TABLE 25.3 Life expectancies from different sources, South African males 2001–2007*

| *x* | Generalized Growth Balance | Princeton West | Iterative estimates | Delta fixed to GGB estimate |
|---|---|---|---|---|
| 65 | 11.8 | 9.45 | 11.6 | 11.7 |
| 70 | 9.4 | 7.37 | 9.3 | 9.4 |
| 75 | 7.4 | 5.55 | 7.3 | 7.4 |
| 80 | 5.7 | 4.06 | 5.6 | 5.7 |
| 85 | 4.4 | 2.90 | 4.3 | 4.4 |

***Step 4: Estimate the number of people who turned x, and the number aged x to x + 4 last birthday, from the reported deaths***

The number of people who turned $x$ during the period between the two censuses as estimated from the numbers of deaths using an open interval of 85+, growth rates from column 6 of Table 25.2 and the estimate of life expectancy at age 85 of 4.347 given in the fifth column of Table 25.3, is as shown in column 7 of Table 25.2. For example, the estimate of the number of people who turned 80 in the period between the censuses is calculated as follows, using the growth rate for the population in open interval 85+ of 0.0638 and the growth rate for the population aged between 80 and 85 of –0.0148, is

$$\hat{N}_{85} = 46196\begin{pmatrix}\exp(0.0638\times 4.347)\\ -(0.0638\times 4.347)^2/6\end{pmatrix} = 61475$$

$$\hat{N}_{80} = 61475\exp(5\times(-0.0148)) + 48163\exp(2.5\times(-0.0148)) = 103519.$$

The number of people aged $x$ to $x+4$ last birthdays during the period between the two censuses, estimated from the reported deaths is given in column 8 of Table 25.2. For example, the number who turned 20 to 24 last birthday is calculated as follows:

$${}_5\hat{N}_{20} = 2.5(2289458.6 + 2059216.1) = 10871687.$$

***Step 5: Estimate the number of people aged x to x + 4 last birthday during the period between the censuses, from the census populations***

The number of people aged between $x$ and $x+5$ during the period between the two censuses appears in column 2 of Table 25.4 and is calculated for the 20 to 24 age group, for example, using the populations given in columns 2 and 3 of Table 25.2 and the time between the two censuses, as follows:

$${}_5N_{20} = 5.3540689(2099417\times 2362519)^{\frac{1}{2}} = 11923972.$$

*TABLE 25.4 The number aged x to x + 4 last birthday estimated from the census population and the ratios of the estimates derived from the numbers of deaths to this, South African males, 2001–2007*

| Age | Obs ${}_5N_x$ | $c$: ${}_5N_x$ | $c$: ${}_{A-x}N_x$ |
|---|---|---|---|
| 0 | 12,636,377 | | |
| 5 | 13,341,976 | 0.8496 | 0.8981 |
| 10 | 13,307,209 | 0.8571 | 0.9050 |
| 20 | 11,923,972 | 0.9118 | 0.9231 |
| 25 | 10,521,174 | 0.9364 | 0.9256 |
| 30 | 9,259,118 | 0.9212 | 0.9230 |
| 35 | 7,998,828 | 0.8943 | 0.9235 |
| 40 | 6,798,761 | 0.9176 | 0.9322 |
| 45 | 5,534,858 | 0.9533 | 0.9371 |
| 50 | 4,426,301 | 0.9586 | 0.9310 |
| 55 | 3,350,250 | 0.9525 | 0.9191 |
| 60 | 2,503,746 | 0.9316 | 0.9028 |
| 65 | 1,856,232 | 0.8954 | 0.8865 |
| 70 | 1,270,220 | 0.8713 | 0.8799 |
| 75 | 798,803 | 0.9037 | 0.8885 |
| 80 | 477,921 | 0.8631 | 0.8631 |

***Step 6: Calculate the ratios of the estimates derived from deaths to those derived from the census populations***

The ratios of the numbers of people aged between $x$ and $x+5$ during the period between the censuses estimated from the reported deaths (column 8 of Table 25.2) to those estimated from the censuses (column 2 of Table 25.4) are given in columns 3 and 4 of Table 25.4. Examples of these calculations for age 65 are as follows:

$$\frac{{}_5\hat{N}_{65}}{{}_5N_{65}} = \frac{1662114}{1856232} = 0.8954$$

$$\frac{{}_{20}\hat{N}_{65}}{{}_{20}N_{65}} = \frac{1662114+1106744+721856+412486}{1856232+1270220+798803+477921} = 0.8865.$$

***Step 7: Estimate the completeness of reporting of deaths***

Setting delta to the intercept of the application of the Generalized Growth Balance method produces a series of ratios which, although reasonably level, appears to fall off with increasing age from about age 50 (see from Figure 25.1). Thus for this example *Solver* was not used to estimate delta.

Completeness was estimated from the ratios in the age range 25 to 64. This was done to avoid, to some extent, biasing the estimate downwards due to the falling off of the ratios at the extreme ages although the method of

determining the estimate is fairly robust to fluctuations at individual ages. This produced an estimate of completeness of 94 per cent as follows:

$$c = 0.5 \times 0.9340 + 0.25(0.9203 + 0.9527)$$
$$= 0.9353 = 94\%.$$

where 0.9340 is the median, 0.9203 the 25th percentile and 0.9527 the 75th percentile of the ratios in column 3 of Table 25.3 between ages 25 and 65.

***Step 8: Estimate mortality rates adjusted for incompleteness of reporting of deaths***

The adjusted population as at the first census date which appears in column 2 of Table 25.5 is the enumerated population given in column 2 of Table 25.2 multiplied by exp(–(–0.00467) × 5.3541) since delta is less than 0. For example the adjusted population for age 20 is

$$2099417 \times \exp\left(-\left(-0.00467 \times 5.3541\right)\right) = 2152629.$$

The adjusted population at the second census date which appears in column 3 of Table 25.5 is the enumerated population given in column 3 of Table 25.2 since delta is less than 0.

Next the deaths are adjusted for incompleteness by dividing the number of reported deaths in each age group shown in column 4 of Table 25.2 by the estimate of completeness. These numbers are shown in column 4 of Table 25.5. For example, for age 20 the number is derived from the number of reported deaths, 54 960, as follows:

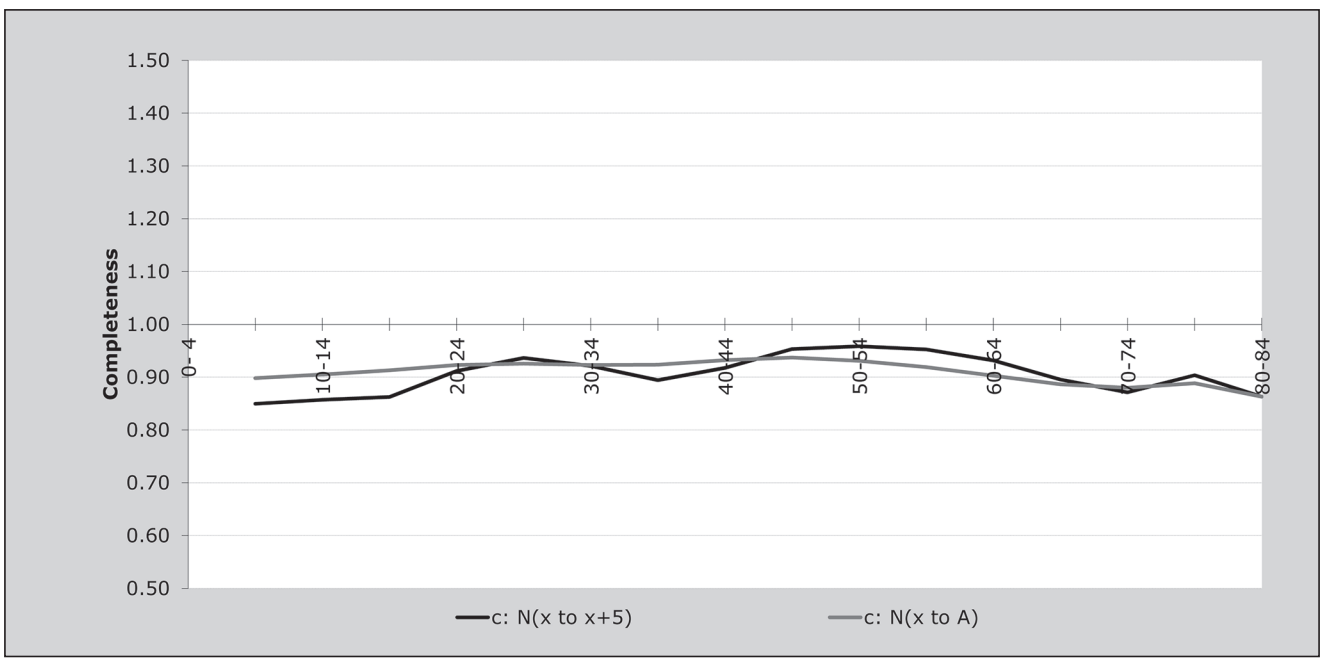
$$\frac{54960}{0.9353} = 58764\,.$$

As it appears that completeness may have declined at the older ages, the option to use age-specific completeness above age 65 is chosen. Thus, for example, the number of deaths between 70 and 75 corrected for incompleteness is calculated as follows:

$$\frac{73036}{0.8713} = 83824\,.$$

The adjusted person-years of life lived (column 5 of Table 25.5) is the geometric average of the populations in columns 2 and 3 of Table 25.5 multiplied by the length (in years) of the period between the censuses, which in this case is 5.3541 years. For age 20 this is

***FIGURE 25.1*** *Completeness by age of registered deaths, South African males, 2001–2007*



$$PYL(20,5) = 5.354069\left(2152629 \times 2362519\right)^{1/2}$$
$$= 12074140.$$

The mortality rates adjusted for incompleteness of reporting of deaths (column 6 of Table 25.5) are derived by dividing the adjusted deaths by the adjusted person-years of life lived. For example, for the 20–24 age group the adjusted rate is calculated as follows:

$$\frac{58764}{12074140} = 0.0049.$$

***Step 9: Smooth using relational logit model life table***

Estimates of probabilities of people aged $x$ dying in the next 5 years, ${}_5q_x$, estimated from the adjusted rates of mortality which appear in column 6 of Table 25.5, are shown in the second column of Table 25.6. For example, the probability of a 20-year old woman dying before reaching age 25 is calculated as follows:

$${}_5q_{20} = \frac{5 \times 0.0049}{1 + 2.5 \times 0.0049} = 0.0240.$$

The life table proportions of five-year olds alive at age $x+5$, estimated from the proportion alive at age $x$ using these values, appear in column 3 of Table 25.6. For example the proportion alive at age 25 is calculated as follows:

$$l_{25} = 0.9796\left(1 - 0.0240\right) = 0.9560.$$

The logit transformations of the proportions surviving appear in column 4 of Table 25.6. For example, the logit transformation of the $l_{20}$ is calculated as follows:

$$\gamma_{20} = 0.5 \ln\left(\frac{1 - 0.9796}{0.9796}\right) = -1.9350.$$

The logit transformation of the conditional life table for males based on the AIDS life table with $e_0$=50 in column 5 of Table 25.6 appears in column 6 of Table 25.6. As can be seen from Figure 25.2, the AIDS model does not fit the data particularly well, but fits better than any table which does not reflect the impact of HIV on mortality.

The coefficients, $\alpha$ and $\beta$ are determined as the intercept and slope of the straight line fitted to the logit transformations in columns 4 and 6 of Table 25.6 over the range of ages chosen by the user (45 and 80 in this example), namely 0.1928 and 1.2008 respectively.

These coefficients are then applied to the logit transformation of the conditional model life table to produce the

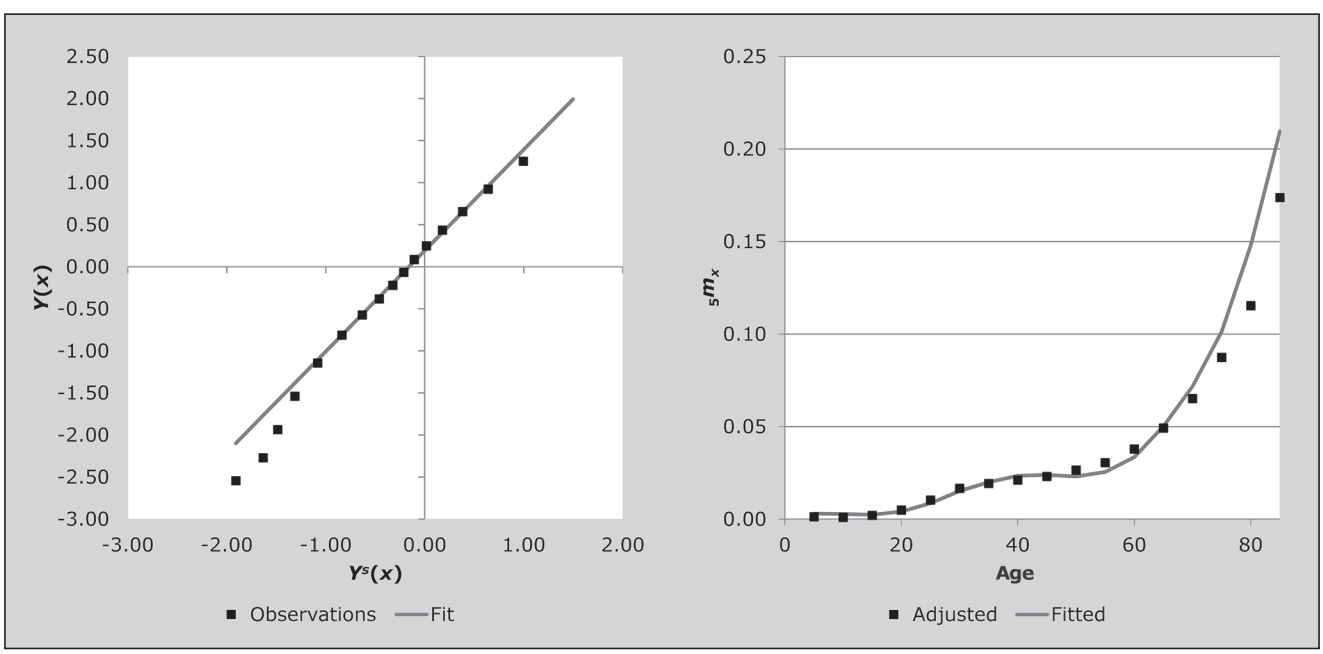
*TABLE 25.5 Calculation of adjusted mortality rates, South African males, 2001–2007*

| Age | Adjusted ${}_5N_x(t_1)$ | Adjusted ${}_5N_x(t_2)$ | Adjusted ${}_5D_x$ | Adjusted $PYL(x,5)$ | Adjusted ${}_5m_x$ |
|---|---|---|---|---|---|
| 0 | | | | | |
| 5 | 2,486,532 | 2,560,642 | 16,644 | 13,510,001 | 0.0012 |
| 10 | 2,582,831 | 2,452,339 | 11,983 | 13,474,797 | 0.0009 |
| 15 | 2,515,334 | 2,553,293 | 27,236 | 13,568,508 | 0.0020 |
| 20 | 2,152,629 | 2,362,519 | 58,764 | 12,074,140 | 0.0049 |
| 25 | 1,947,414 | 2,033,165 | 109,919 | 10,653,675 | 0.0103 |
| 30 | 1,635,041 | 1,875,483 | 155,667 | 9,375,725 | 0.0166 |
| 35 | 1,478,197 | 1,548,185 | 156,001 | 8,099,564 | 0.0193 |
| 40 | 1,265,085 | 1,306,900 | 145,347 | 6,884,383 | 0.0211 |
| 45 | 992,273 | 1,104,294 | 129,387 | 5,604,563 | 0.0231 |
| 50 | 789,134 | 888,042 | 118,852 | 4,482,045 | 0.0265 |
| 55 | 566,403 | 708,812 | 103,560 | 3,392,442 | 0.0305 |
| 60 | 455,861 | 491,871 | 96,156 | 2,535,277 | 0.0379 |
| 65 | 312,561 | 394,305 | 92,518 | 1,879,609 | 0.0492 |
| 70 | 238,500 | 241,976 | 83,824 | 1,286,217 | 0.0652 |
| 75 | 139,925 | 163,112 | 70,679 | 808,863 | 0.0874 |
| 80 | 93,159 | 87,698 | 55,803 | 483,940 | 0.1153 |
| 85 | 47,084 | 70,299 | 53,524 | 308,032 | 0.1738 |

*TABLE 25.6 Calculation of smoothed mortality rates using a relational logit model life table, South African males, 2001–2007*

| Age | ${}_5q_x$ | $l_x/l_5$ | Obs. $Y(x)$ | AIDS Cdn. $l^s(x)$ | Cdn. $Y^s(x)$ | Fitted $Y(x)$ | Fitted $l(x)$ | $T(x)$ | $e(x)$ | Smooth ${}_5m_x$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | | | | | | | | | |
| 5 | 0.0061 | 1 | | 1.0000 | | | 1 | 51.206 | 51.2 | 0.0030 |
| 10 | 0.0044 | 0.9939 | –2.5433 | 0.9785 | –1.9081 | –2.0984 | 0.9852 | 46.243 | 46.9 | 0.0028 |
| 15 | 0.0100 | 0.9894 | –2.2705 | 0.9632 | –1.6326 | –1.7676 | 0.9717 | 41.351 | 42.6 | 0.0024 |
| 20 | 0.0240 | 0.9796 | –1.9350 | 0.9512 | –1.4853 | –1.5907 | 0.9601 | 36.521 | 38.0 | 0.0041 |
| 25 | 0.0503 | 0.9560 | –1.5395 | 0.9324 | –1.3120 | –1.3827 | 0.9408 | 31.769 | 33.8 | 0.0086 |
| 30 | 0.0797 | 0.9079 | –1.1444 | 0.8969 | –1.0818 | –1.1062 | 0.9014 | 27.164 | 30.1 | 0.0152 |
| 35 | 0.0919 | 0.8356 | –0.8128 | 0.8420 | –0.8365 | –0.8116 | 0.8352 | 22.822 | 27.3 | 0.0200 |
| 40 | 0.1003 | 0.7588 | –0.5731 | 0.7794 | –0.6311 | –0.5650 | 0.7559 | 18.845 | 24.9 | 0.0235 |
| 45 | 0.1091 | 0.6827 | –0.3831 | 0.7148 | –0.4593 | –0.3588 | 0.6721 | 15.275 | 22.7 | 0.0239 |
| 50 | 0.1243 | 0.6082 | –0.2199 | 0.6560 | –0.3228 | –0.1948 | 0.5962 | 12.104 | 20.3 | 0.0230 |
| 55 | 0.1418 | 0.5326 | –0.0653 | 0.6048 | –0.2127 | –0.0626 | 0.5313 | 9.285 | 17.5 | 0.0255 |
| 60 | 0.1732 | 0.4571 | 0.0861 | 0.5530 | –0.1064 | 0.0650 | 0.4676 | 6.788 | 14.5 | 0.0335 |
| 65 | 0.2191 | 0.3779 | 0.2493 | 0.4918 | 0.0163 | 0.2124 | 0.3954 | 4.631 | 11.7 | 0.0502 |
| 70 | 0.2802 | 0.2951 | 0.4354 | 0.4119 | 0.1781 | 0.4066 | 0.3072 | 2.874 | 9.4 | 0.0718 |
| 75 | 0.3586 | 0.2124 | 0.6553 | 0.3178 | 0.3819 | 0.6513 | 0.2137 | 1.572 | 7.4 | 0.1013 |
| 80 | 0.4475 | 0.1362 | 0.9235 | 0.2173 | 0.6408 | 0.9622 | 0.1274 | 0.719 | 5.6 | 0.1480 |
| 85 | #N/A | 0.0753 | 1.2542 | 0.1201 | 0.9959 | 1.3887 | 0.0586 | 0.255 | 4.3 | 0.2097 |

*FIGURE 25.2 Observed logits and adjusted mortality rates against expected derived from the male AIDS model life table, South African males, 2001–2007*



fitted logits in column 7 of Table 25.6. Thus, for example, the fitted logit at age 20 is calculated as follows:

$$\gamma_{20}^{fitted} = 0.1928 + 1.2008 \times (-1.4853) = -1.5907.$$

These values are then used to produce the fitted life table in column 8 of Table 25.6. For example the value at age 20 is calculated as follows:

$$l_{20}^{fitted} = \frac{1}{1 + \exp(2 \times (-1.5907))} = 0.9601.$$

The conditional years of life lived, $T_x$, which appear in column 9 of Table 25.6, are then calculated from the fitted life table. These numbers are then used to produce the smoothed mortality rates which appear in column 10 of Table 25.6. For example, for age 80,

$$T_{80} = 0.255 + \frac{5}{2}(0.1274 + 0.0586) = 0.719$$

$$_5m_{80}^{fitted} = \frac{0.1274 - 0.0586}{0.719 - 0.2655} = 0.1480.$$

The life expectancies which appear in column 10 of Table 25.6 are the numbers in column 9 divided by the numbers in column 8. For example, the life expectancy at age 65 is

$$\frac{4.631}{0.3954} = 11.71.$$

## DIAGNOSTICS, ANALYSIS AND INTERPRETATION

### Checks and validation

The estimate of completeness is 94 per cent. The first check on this result is a comparison with the results for the opposite sex. For example, applying the same method as described above for men to the data for women during the same period (in the *SEG_South Africa_female*s workbook) gives an estimate of completeness of 93 per cent. Past research (Dorrington, Moultrie and Timæus 2004) leads to the expectation that the estimates should be similar, so the results are sufficiently close as to validate the estimates.

A second check on the results is to compare them with the result from the Generalized Growth Balance method (in the *GGB_South Africa_males* workbook), which estimated the completeness of death reporting over the age range 5 to 84 to be 92 per cent. This again supports the result.

A third check is to compare estimates of various key indicators of mortality with those from other sources, such as previous estimates for the country or the *World Population Prospects* (UN Population Division 2011). The estimate of $_{45}q_{15}$ from the observed mortality rates after adjusting for incompleteness is 51.9 per cent, while the estimate of $_{45}q_{15}$ from the WPP for the period 2000–2005 is 52.9 per cent, again suggesting no reason to question the results.

As a matter of interest, application of the Preston and Coale method to these data (estimating the population in the middle of the period as that average of the two survey populations) provides an estimate of completeness, using the same age range, of 84 per cent. Increasing the minimum age of range of the data used to determine delta to 35 increases the estimate to 86 per cent, still somewhat lower than the estimate of 94 per cent produced above.

### Interpretation

Inspection of the estimates of completeness (Figure 25.1) suggests that the completeness of death reporting appears to fall steadily with age from about age 55, which is consistent with migrant workers retiring from urban areas to rural areas, where completeness of registration was lower. Since these estimates were produced using the estimate of delta produced by application of the Generalized Growth Balance method and it is entirely plausible that people have retired from urban to rural areas, reducing delta to produce a more level set of estimates is inappropriate.

Since migration has been taken into account, the falloff in estimates of completeness at the younger ages is probably due to the opposite of what is happening at the older ages, namely, young people moving from rural to urban areas to find employment. It would be wrong, therefore, to allow estimates at these ages to influence the estimate of overall completeness unduly.

The lack of smoothness in the series of the ratios of $_5N_x$ is determined by the estimates of the population from the census and survey, and not from the deaths. Thus, the erratic nature of this series is probably indicative of errors due to relative undercounting in particular age groups and/or age misreporting in the census or survey population estimates.

## METHOD-SPECIFIC ISSUES WITH INTERPRETATION

### Source of reported deaths

Generally there are two sorts of problems with the death data: those that lead to under/over coverage that is constant by age, which is precisely what the method is intended to address, and those which lead to differential coverage by age, which can distort the estimates. Although the general approach remains essentially the same irrespective of the source of the death data, different sources of death data are prone to different biases which might impact on the interpretation of the results. These are illustrated by way of particular examples, but, in general terms, the analyst needs to look out for the following biases in the death data.

#### *1) Vital registration*

If the proportionate split of the population between urban and rural (or appropriate proxies) areas differs significantly by age and the completeness of reporting of deaths in urban areas is significantly higher than it is in rural areas, then the assumption that completeness is independent of age is likely to be violated by a falling off of completeness with age at ages over 50 if a proportion of people move from urban to rural areas on retirement. If ignored, this violation is likely to lead to an underestimate of the average level of completeness.

#### *2) Deaths reported by households*

The data are subject to four potential problems:

- If a significant proportion of households dissolve on the death of a key person (e.g. the sole breadwinner), then the deaths of such people go unreported, leading to a violation of the assumption that completeness is invariant with age. If a significant proportion of deaths in some age groups are of individuals who do not live in private households (for example, they live in homes for the elderly), the breach of the assumption could be even more severe. However, this is not an issue in most developing countries.
- In situations where young adults leave the home they grew up in to work in urban areas, it is possible that they are regarded as being members of more than one household (or of neither household) and their deaths could be reported more than once (or not at all), again leading to a violation of the assumption of constant reporting of deaths by age. In this case, one can limit the impact by ignoring the data below a specific age in determining completeness.
- Reference period error: Since there is often confusion about the exact period for which deaths are to be reported, not to mention uncertainty about exact dates of death, it is possible for there to be overall under- or over-reporting of deaths. Provided one can assume that this is independent of the age of the deceased, this distortion will be accounted for in the estimate of completeness and is not a problem for estimating mortality rates.
- The reference period covers a small proportion of the intercensal period, for example, the common situation in which households report on deaths for the year preceding the census. Not only might such a short period result in significant random fluctuation, but in addition one does not have an estimate of the population at the start of this reference period. How one might deal with this is illustrated in the examples given. Essentially, if one has, in addition, deaths reported by households at the first census, one can use the two sets of data on deaths to estimate the number of deaths during the intercensal period, as was discussed above. However, since the question asking households to report on deaths in the previous year was used relatively seldom before the 2010 round of censuses, one may only have the single set of data on deaths. In this case, provided there are no reasons for assuming that the age pattern of mortality has changed rapidly over the period, it is recommended that one calculates the age-specific death rates for the year and applies these to the person-years of life lived for the interval to get an estimate of deaths for the period. If there are reasons for suspecting that mortality has changed rapidly, for example due to HIV/AIDS, then this adaptation is likely to underestimate or overestimate the mortality and the use of death distribution methods is not recommended.

#### *3) Deaths recorded in health facilities*

Little is known about how well this source of data works. However, it can be expected that completeness would depend on the distribution of health services from which the data have been gathered, and in many developing countries such services are likely to be concentrated in urban areas. So, again, if the proportion of the population living in urban rather than rural areas varies with age, then completeness cannot be assumed to be independent of age. It is also possible that certain causes will predominate in facilities, and if these causes are significant and age-related, this could lead to a further violation of the assumption of constant completeness by age.

In all such cases, one should avoid the temptation of adjusting delta to produce a level sequence of the ratios, and ensure that the estimate of *c* is determined over an age range which excludes the ages where distortions exist.

**General diagnostic interpretation**

In practice both the sequences of ${}_5\hat{N}_x \big/ {}_5N_x$ and ${}_A\hat{N}_x \big/ {}_AN_x$ are affected by violations of the assumptions. However, part of the power of this technique is that most of the typical violations of assumptions produce fairly distinctive characteristic deviations from the expected horizontal plot and in certain circumstances these patterns are interpretable. The following are examples:

- *Incorrect estimate of relative coverage of the censuses*: If $\delta$ is too high the sequences fall nearly linearly with increasing age towards the underlying value of completeness and vice versa, as can be concluded from inspection of Equation 1 below. The effect is greater for ${}_5\hat{N}_x \big/ {}_5N_x$ than for ${}_A\hat{N}_x \big/ {}_AN_x$.
- *Exaggeration of reported age*: Typically, relatives reporting deaths exaggerate the person's age at death more than living individuals reporting their own ages. This produces rising sequences of points which are imperceptible up to the age at which exaggeration begins, followed by a sharp upward curve thereafter. Again, it can be seen from inspection of Equation 1 below, that age exaggeration leads to an increase in the number of deaths in the older age categories. In addition, transfers within an age category lead to those deaths being multiplied by a larger exponential term, although this effect is far smaller. Although such a pattern would also be produced by rising completeness in death registration with age beyond a certain age there appears to be no evidence of this in practice (Preston, Coale, Trussell *et al.* 1980).
- *Age misstatement in the population estimates and age-specific miscounting*: This is exhibited by an erratic sequence of the ratios over the age span. Since ${}_5\hat{N}_x$ is cumulative in form, it tends to follow the age distribution of the population quite closely and hence, if there are zigzags, it is likely that the peaks are associated with inflated population estimates and the troughs with deflated ones. If these fluctuations are independent of age they will not distort the estimate of completeness particularly. However, if they are systematic, for example due to unaccounted for migration beneath a certain age, it may be desirable not to include these points in estimating the completeness.

## EXAMPLES USING DEATHS REPORTED BY HOUSEHOLDS IN A CENSUS/SURVEY

The examples below use the same data as used in the *SEG_South Africa_males* and *SEG_South Africa_females* workbooks with the exception that instead of using the vital registration as the source of the death data, deaths are estimated from deaths reported by households in the 2001 Census and the 2007 Community survey as having occurred in the year preceding the census/survey. These numbers are given in Table 25.7.

The numbers of deaths occurring between the date of the Census (midnight between 9 and 10 October 2001) and the survey (assumed to be midnight between 14 and 15 February 2007) are estimated using the *Estimating deaths_South Africa_males_hhd* and the *Estimating deaths_South Africa_females_hhd* workbooks.

Applying the Synthetic Extinct Generations method to these data for males using the estimate of relative incompleteness of census coverage (delta) derived from the application of the Generalized Growth Balance method to these data, *SEG_South Africa_males_hhd*, suggests that these estimates of the number of deaths are more or less as

*TABLE 25.7 Deaths reported by households to have occurred in the year preceding census/survey, South Africa*

| | 2001 Census | | 2007 Community Survey | |
|---|---|---|---|---|
| Age | Males | Females | Males | Females |
| 0–4 | 35,873 | 32,096 | 48,322 | 44,418 |
| 5–9 | 3,868 | 3,155 | 4,505 | 5,216 |
| 10–14 | 2,590 | 2,284 | 3,442 | 3,259 |
| 15–19 | 5,628 | 5,122 | 8,246 | 7,878 |
| 20–24 | 10,976 | 13,246 | 16,360 | 21,702 |
| 25–29 | 17,787 | 19,727 | 27,551 | 35,840 |
| 30–34 | 20,038 | 18,292 | 34,832 | 42,576 |
| 35–39 | 19,816 | 15,521 | 38,061 | 34,809 |
| 40–44 | 17,417 | 12,124 | 33,604 | 28,823 |
| 45–49 | 15,840 | 10,105 | 27,829 | 20,973 |
| 50–54 | 15,077 | 9,144 | 28,223 | 18,891 |
| 55–59 | 12,781 | 7,755 | 22,868 | 13,118 |
| 60–64 | 13,428 | 10,367 | 18,775 | 14,912 |
| 65–69 | 11,820 | 10,195 | 17,532 | 14,298 |
| 70–74 | 11,885 | 10,809 | 14,879 | 14,645 |
| 75–79 | 8,794 | 8,393 | 12,966 | 14,151 |
| 80–84 | 7,484 | 9,371 | 9,204 | 12,063 |
| 85+ | 7,115 | 12,389 | 11,735 | 18,178 |

completely reported as the vital registration. However these data estimate ${}_{45}q_{15}$ at 53.9 per cent, which although similar, is slightly higher than the estimate produced using registered deaths. Applying the Synthetic Extinct Generations method to these data for females, *SEG_South Africa_females_hhd*. suggests that the deaths of women reported by households are far less complete than the registered deaths and estimates ${}_{45}q_{15}$ at 49.3 per cent. This is higher (and less plausible relative to the probability for males) than the 42 per cent produced using registered deaths.

The reason for the much poorer performance of the method applied to deaths of women reported by households can be seen by a comparison of the estimated numbers of deaths for the period derived from deaths reported by households to the numbers expected after correcting the vital registration for incompleteness of reporting, as shown in Table 25.8. From this we see that there is a significant decline in completeness of reporting of deaths of women by households with age from age 55, probably as the result disintegration of households on the death of these women, usually because these households were headed by the women who died.

There is also evidence of over-reporting of deaths below age 30 for males and 25 for females.

In the situation where only the most recent census asked about deaths in the previous year, the number of deaths in each age group between the times of the 2001 Census and the 2007 Community Survey using only the deaths reported by households in the 2007 Community Survey are estimated as follows:

$$\frac{{}_5D_x(t_2)}{{}_5N_x(t_2)}(t_2-t_1)\left({}_5N_x(t_1)\times {}_5N_x(t_2)\right)^{1/2}.$$

Applying the method to these estimates of the deaths produce estimates of ${}_{45}q_{15}$ of 58.1 per cent for males and 55.6 per cent for females. Unlike the previous estimates, these are estimates of mortality in the year preceding the second census/survey and therefore might be expected to be higher than those for the whole period, since mortality has been increasing over the period due to HIV/AIDS. However, as might also be expected, deriving an estimate from a single year of deaths (derived, in addition, in this case from a relatively small sample survey) produces far less reliable estimates, particularly in the case (for these data) of females. Alternative estimates (Bradshaw, Dorrington and Laubscher 2012) suggest that for 2006 the correct probabilities should be closer to 55 per cent for males and 45 per cent for females.

| | Males | | | Females | | |
|---|---|---|---|---|---|---|
| Age | Reported | Expected | Ratio | Reported | Expected | Ratio |
| 0–4 | | | | | | |
| 5–9 | 22,683 | 16,979 | 134% | 22,995 | 14,575 | 158% |
| 10–14 | 16,462 | 12,224 | 135% | 15,173 | 10,349 | 147% |
| 15–19 | 38,013 | 27,784 | 137% | 35,666 | 26,874 | 133% |
| 20–24 | 74,934 | 59,946 | 125% | 95,993 | 84,611 | 113% |
| 25–29 | 124,403 | 112,129 | 111% | 152,718 | 154,437 | 99% |
| 30–34 | 150,792 | 158,796 | 95% | 166,488 | 170,680 | 98% |
| 35–39 | 159,016 | 159,137 | 100% | 137,837 | 141,399 | 97% |
| 40–44 | 140,172 | 148,269 | 95% | 111,910 | 115,746 | 97% |
| 45–49 | 120,016 | 131,988 | 91% | 85,284 | 93,408 | 91% |
| 50–54 | 118,989 | 121,242 | 98% | 76,941 | 81,793 | 94% |
| 55–59 | 97,977 | 105,641 | 93% | 57,353 | 72,131 | 80% |
| 60–64 | 88,088 | 98,089 | 90% | 69,220 | 78,877 | 88% |
| 65–69 | 80,451 | 90,359 | 89% | 67,007 | 86,099 | 78% |
| 70–74 | 72,827 | 79,663 | 91% | 69,536 | 93,404 | 74% |
| 75–79 | 59,632 | 69,665 | 86% | 61,942 | 88,314 | 70% |
| 80–84 | 45,365 | 52,533 | 86% | 58,410 | 77,084 | 76% |
| 85+ | 51,779 | 50,387 | 103% | 83,753 | 108,002 | 78% |

***TABLE 25.8*** *Ratio of estimates of deaths derived from deaths reported by households to the expected numbers of deaths, South Africa*

## DETAILED DESCRIPTION OF METHOD

### Mathematical exposition

The basic idea of the Synthetic Extinct Generations method was originally proposed by Vincent (1951), namely that the number of persons at a particular age at a point in time must equal the total number of deaths arising from this population from that time until the last survivor has died. This method is a generalization of the method proposed by Preston, Coale, Trussell *et al.* (1980), which required that the population be stable. Thus, for a cohort followed prospectively,

$$N(a,t) = \int_0^{\varpi - a} D(a+s, t+s)ds\,,$$

where $N(a,t)$ represents the number of people in the population at time $t$, aged $a$ exactly and $D(a+s, t+s)ds$ represents the number of deaths at time $t+s+ds$, aged $a+s$ to $a+s+ds$.

These numbers of deaths in turn can be estimated from the number of deaths recorded in a given interval, on the assumption that mortality rates remain constant over that interval, by noting that deaths at any particular age will grow at the accumulated population growth rate at that age from time $t$ to $t+s$. That is,

$$D(a+s,t+s) = D(a+s,t)e^{\int_0^s r(a+s,t+z)dz},$$

where $r(a+s, t+z)dz$ represents the annual growth rate of the population aged $a+s$ in the time interval $t+z$ to $t+z+dz$. If mortality is unchanging over time, it can be shown that $r(a+s, t+z) = r(a+s-z, t)$ and hence that

$$N(a,t) = \int_0^{\varpi - a} D(a+s,t)e^{\int_0^s r(a+s-z,t)dz} ds\,,$$

which can be rewritten as

$$N(a,t) = \int_a^{\varpi} D(x,t)e^{\int_a^x r(y,t)dy} dx\,. \qquad (1)$$

Thus, it is possible to derive an estimate of the population at each age based on the age-specific numbers of deaths over a particular interval and the age-specific population growth rates. Comparison of these estimates of the population to the estimates derived from censuses gives an estimate of completeness of the deaths relative to that population.

### Implementation of the method in practice

Since the data one typically works with are subject to age misstatement, in practice one usually works with data grouped into five-year age groups.

Thus, assume that in practice one has data on the following: the number of reported deaths over a number of years, from times $t_1$ to $t_2$, in five-year age groups, ${}_5D_x^r$, up to an open interval at age $A$, ${}_\infty D_A^r$; and the number of people in the population at each of times $t_1$ and $t_2$, in the same age groups reported by the censuses, ${}_5N_x^r(t)$ up to ${}_\infty N_A^r(t)$ (where $t$ = either $t_1$ or $t_2$). These data can then be used to apply the method by computing ${}_5D_x^r$ and ${}_\infty N_x^r$, and approximating $N_x^r$ by $(t_2 - t_1)\left({}_5N_{x-5}^r(t_1) \times {}_5N_x^r(t_2)\right)^{1/2} \Big/ 5$ and $\int_{t_1}^{t_2} {}_\infty N_x^r(t)\,dt$ by $(t_2 - t_1)\left({}_\infty N_x^r(t_1) \times {}_\infty N_x^r(t_2)\right)^{1/2}$.

As suggested by Bennett and Horiuchi (1981), a computational form of Equation 1 can be derived, namely that

$$N_x = N_{x+5}\exp\left[5 \times {}_5r_x\right] + {}_5D_x \exp\left[2.5 \times {}_5r_x\right], \qquad (2)$$

where $N_x$ represents the number of people in the population who turned $x$ between the census at time $t_1$ and the census at time $t_2$, and ${}_5r_x$ represents the average annual growth rate of the population aged $x$ to $x+4$ last birthday between times $t_1$ and $t_2$.

Bennett and Horiuchi (1981, 1984) suggest using the age group-specific growth rates to improve both the $\hat{N}_x$ at the older ages and $\hat{N}_A$, where $A$ is the age at the start of the open-ended age interval. They suggest calculating $\hat{N}_A$ as follows:

$$\hat{N}_A = {}_\infty D_A^r \left[\exp\left({}_\infty r_A \times e_A\right) - \left({}_\infty r_A \times e_A\right)^2 / 6\right],$$

where ${}_\infty D_A^r$ represents the reported deaths aged $A$ and older, and $e_A$ represents life expectancy at age $A$.

They also suggest that in order to allow for the curvature, particularly at the older ages, Equation 2 could be modified as follows:

$$\hat{N}_x = \hat{N}_{x+5}\exp\left[5 \times {}_5r_x\right] + {}_5\gamma_x \times {}_5D_x^r \exp\left[2.5 \times {}_5r_x\right],$$

where

$${}_5\gamma_x = 1.00 - 2.26 \times {}_5r_x \frac{{}_5D_x^r}{{}_5N_x} + 0.218 \times 5r_x - 0.826\left({}_5r_x\right)^2.$$

In addition to this, they suggest that over the age of 60 the ${}_5\hat{N}_x$ be approximated by "imposing a stable population curve over the five-year span and then determining the area under the curve accordingly" (Bennett and Horiuchi 1981: 210). However, in practice the data are rarely accurate enough to warrant such a refinement, and it seldom makes much difference to the estimate of completeness.

In practice, in order to correct, to some extent, for the effects of digit preference in age reporting and also to make the death data consistent with population data for conventional five-year age groups, it is usual to compute ${}_5\hat{N}_x = 2.5(\hat{N}_x + \hat{N}_{x+5})$.

Furthermore, since the sequence of ${}_5\hat{N}_x \big/ {}_5N_x$ ratios (or even ${}_{10}\hat{N}_{x-5} \big/ {}_{10}N_{x-5}$ as suggested by Bennett and Horiuchi (1981)) is usually still somewhat erratic because of age misreporting and differential omission of persons in particular age spans, it is usual to assume that the percentage reported is roughly constant with respect to age for ages greater than, say, 10. One estimates this fixed proportion, $c$, by either the mean or median of the values of ${}_5\hat{N}_x \big/ {}_5N_x$ over a representative span of ages (after, if necessary, correcting the age group-specific growth rates for the differential completeness of the two censuses). Allowing for differential completeness of the two censuses is achieved by adding a constant factor, $\delta$, to the age group-specific growth rates derived from the reported population numbers to produce a 'flat sequence' of ${}_5\hat{N}_x \big/ {}_5N_x$.

To see this, suppose

$$ {}_\infty N_x^r(t_1) = k_1 {}_\infty N_x(t_1) \text{ and } {}_\infty N_x^r(t_2) = k_2 {}_\infty N_x(t_2). $$

Then

$$ {}_5r_x^r = \ln\left( \frac{{}_\infty N_x^r(t_2)}{{}_\infty N_x^r(t_1)} \right) \Big/ t = \ln\left( \frac{k_2 \, {}_\infty N_x(t_2)}{k_1 \, {}_\infty N_x(t_1)} \right) \Big/ t $$

$$ = {}_5r_x + \ln\left( \frac{k_2}{k_1} \right) \Big/ t = {}_5r_x - \delta, $$

where

$$ \delta = \ln\left( \frac{k_1}{k_2} \right) \Big/ t \, . $$

Thus $1 - \delta \cdot t$, where $t$ is the length of the intercensal period, gives an indication of the differential completeness between the two estimates of the population used to estimate the age group specific growth rates.

The sequence of ${}_5\hat{N}_x \big/ {}_5N_x$ values is usually plotted together with that of ${}_A\hat{N}_x \big/ {}_AN_x$, where ${}_A\hat{N}_x = \sum_{a=x}^{A-5} {}_5\hat{N}_a$. The latter ratio tends to be more stable and assists with interpretation of the data.

**Extensions**

If the ages were recorded accurately and the assumption of constant census coverage by age held (not a very likely situation), then the method could be adapted to deal with the situation where completeness of reporting of the deaths was constant only for a limited age range ($x$ to $x+n$) by applying a truncated version of the method which eliminates from consideration deaths and population aged $x+n$ and older. This adaptation could, for example, be applied to vital registration data where completeness might fall off above retirement age if people retired from urban to rural areas. It could also be applied where deaths reported by households where household might disintegrate on the death of the last adult. However, unless $x+n$ is high this method is unlikely to be very robust. Analogous adaptation of the Generalized Growth Balance method is easier and seems to be a little more robust.

## FURTHER READING AND REFERENCES

Analysis of the sensitivity of the method to common data errors and violation of the assumptions is fairly limited. However, the reader is referred to Hill, You and Choi (2009) for an analysis of the assumptions underlying the death distribution methods in the absence of HIV and to Dorrington and Timæus (2008) for an analysis in a population experiencing significant HIV. Murray, Rajaratnam, Marcus *et al.* (2010), in contrast, used stochastic simulations to assess these methods. However, to date their work has had very limited impact on the use of Death Distribution methods, possibly because their description of their simulations is short on detail and because their assessment is based on perhaps unrealistically high migration.


Bennett NG and S Horiuchi. 1981. "Estimating the completeness of death registration in a closed population", *Population Index* **47**(2):207–221. doi: 10.2307/2736447

Bennett NG and S Horiuchi. 1984. "Mortality estimation from registered deaths in less developed countries", *Demography* **21**(2):217–233. doi: 10.2307/2061041

Blacker J. 1988. *An Evaluation of the Pakistan Demographic Survey*. Karachi: Pakistan Federal Bureau of Statistics.

Bradshaw D, RE Dorrington and R Laubscher. 2012. *Rapid Mortality Surveillance Report 2012*. Cape Town: South African Medical Research Council. http://www.mrc.ac.za/bod/RapidMortality2011.pdf

Bradshaw D, RE Dorrington and R Laubscher. 2012. *Rapid Mortality Surveillance Report 2012*. Cape Town: South African Medical Research Council. http://www.mrc.ac.za/bod/RapidMortality2011.pdf

Dorrington RE, TA Moultrie and IM Timæus. 2004. *Estimation of Mortality using the South African 2001 Census Data*. Monograph 11. Centre for Actuarial Research, University of Cape Town. http://www.commerce.uct.ac.za/care/Monographs/Monographs/ Mono11.pdf

Dorrington RE and IM Timæus. 2008. "Death distribution methods for estimating adult mortality: sensitivity analysis with simulated data errors, revisited," Paper presented at Population Association of America 2008 Annual Meeting. New Orleans, Louisiana, 17–19 April.

Hill K, D You and Y Choi. 2009. "Death distribution methods for estimating adult mortality: Sensitivity analysis with simulated data error", *Demographic Research* **21**(Article 9):235–254. doi: 10.4054/DemRes.2009.21.9

Murray CJL, JK Rajaratnam, J Marcus, T Laakso and AD Lopez. 2010. "What can we conclude from death registration? Improved methods for evaluating completeness", *PLoS Med* **7**(4):e1000262. doi: 10.1371/journal.pmed.1000262

Preston SH, AJ Coale, J Trussell and M Weinstein. 1980. "Estimating the completeness of reporting of adult deaths in populations that are approximately stable", *Population Index* **46**:179–202. doi: 10.2307/2736122

Timæus IM. 2004. "Impact of HIV on mortality in Southern Africa: Evidence from demographic surveillance," Paper presented at Seminar of the IUSSP Committee "Emerging Health Threats" HIV, Resurgent Infections and Population Change in Africa. Ougadougou, 12–14 February.

UN Population Division. 2011. *World Population Prospects: The 2010 Revision, Volume I: Comprehensive Tables*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/313. http://esa.un.org/unpd/wpp/Documentation/pdf/WPP2010_Volume-I_Comprehensive-Tables.pdf

Vincent P. 1951. "La mortalite des vieillards", *Population* **6**:182–204. doi: 10.2307/1524149

# Chapter 26 Indirect estimation from orphanhood in multiple inquiries

*Ian M Timæus*

## DESCRIPTION OF THE METHOD

The basic orphanhood method estimates the mortality of adult women and men indirectly from data on the survival status of respondents' mothers and fathers respectively. In order to apply the method, censuses and surveys must minimally have included the questions 'Is your mother alive?' and/or 'Is your father alive?'. By allowing for the mean age at which the mothers and fathers have children in the population concerned, it is possible to convert the proportion of persons in each age group with living mothers and living fathers into life table measures of survivorship in adulthood ($l_{25+n}/l_{25}$ for women and $l_{35+n}/l_{35}$ for men).

Once data on orphanhood have been collected in two successive inquiries, it is possible to derive synthetic cohort measures of parental survival for the intervening period and estimate life table measures for this period from them. In particular, if adults aged 15 to 49 have been asked about the survival of their parents, one can estimate survivorship from orphanhood in adulthood, which is to say for synthetic cohorts based at age 20. Synthetic cohort methods can provide estimates of adult mortality for a clearly defined and relatively up-to-date period. This is especially useful in countries experiencing generalized HIV epidemics where the level of adult mortality is likely to have changed abruptly in the last couple of decades. The approach also potentially reduces bias resulting from underreporting of orphanhood by respondents who were orphaned at a young age.

If a supplementary question has been asked in a single inquiry about when dead parents died, one can use this information to reconstruct the proportions of respondents with living mothers and fathers at earlier dates and analyse orphanhood in the intervening periods in the same way as data from multiple inquiries.

One advantage that orphanhood methods have over direct questions about deaths in households is that adult mortality can be estimated in this way in moderately-sized inquiries. In contrast, only censuses or unusually large surveys can yield direct estimates based on deaths in the year before the inquiry that are sufficiently precise to be useful. Moreover, orphanhood methods do not assume that the population is closed to migration. However, the results from them will not be representative for small states or sub-national areas in which a substantial proportion of the population are in-migrants or have emigrated.

## DATA REQUIREMENTS

To estimate the mortality of adult women:

- The proportions of respondents whose mother is alive by five-year age group of respondent at two or more different dates. (Those who did not know or did not declare their mother's survival status should be excluded from the calculations.)
- The number of births in the year before a demographic inquiry tabulated by five-year age group of women giving birth.

To estimate the mortality of adult men:

- The proportion of respondents whose father is alive by five-year age group of respondent at two or more different dates. (Those who did not know or did not declare their father's survival status should be excluded from the calculations.)
- The number of births in the year before a demographic inquiry tabulated by five-year age group of the women giving birth.
- An estimate of the difference between the ages of men and women having children, such as the difference between the median ages of currently married men and women.

These tables should generally be produced separately for

male and female respondents and estimates made from both sets of proportions and for the two sexes combined.

The synthetic cohort method described here estimates adult mortality from orphanhood data supplied by adult respondents, which is to say those aged 15 or more years. While no data on younger age groups are required to produce the synthetic cohort estimates, if they were collected they should usually be entered into the spreadsheet so that they can be used to produce estimates by means of the basic orphanhood method.

If sample or design weights have been provided with the data, remember to apply them in the manner appropriate to your statistical software when deriving the tabulations used as inputs.

## IMPORTANT ASSUMPTIONS

An inherent limitation of the orphanhood method is that data on parents' survival can only be collected from those of their offspring who are alive themselves. The survival of adults who have no living children is unrepresented in the reported proportions of parents alive. Moreover, parents with more than one surviving child are over-represented in comparison to those with exactly one surviving child in proportion to the number of their surviving children. Thus, the method only produces unbiased results if the mortality of parents is unrelated to how many of their children are alive at the time that the data are collected. In general though, the selection bias that arises from breaches in this assumption is small (Palloni, Massagli and Marcotte 1984). In populations affected by generalized HIV epidemics, however, it is likely to be more severe.

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Before starting the analysis, one should check how many respondents stated that they did not know whether their mother or, more commonly, father was alive or failed to answer the questions at all. The response rate on these questions is usually very high and one can simply exclude from the analysis those respondents who either answered 'don't know' or did not answer the question. In effect, this amounts to assuming that the proportion of these respondents' parents that have died is the same as for respondents that answered the question. However, a few surveys have collected sufficiently incomplete data to suggest that non-response bias could be a substantial problem. For example, it is possible that most people who fail to answer the question have dead parents. If this is the case, such unreported orphans could represent a substantial proportion of all orphans, particularly in the younger age groups, producing a substantial downward bias in the final estimates of mortality.

One useful check on the quality of the orphanhood data is to compare the responses of male and female respondents of the same age. One would not expect the proportion of parents that have died to differ significantly between men and women of the same age. If the proportions diverge among older respondents, this could reflect gender differences in patterns of age misreporting or could indicate that the gender that reports fewer dead parents (usually the men) is more likely to lose touch with their families and is assuming wrongly that some parents remain alive who have died.

When two or more sets of data on the survival of parents are available, one should usually estimate mortality from each set of data independently using the basic orphanhood method as well as produce estimates from synthetic cohort data on orphanhood in adulthood in order to compare the three sets of results. The accompanying Excel workbook (see website) produces both basic and synthetic cohort estimates.

## CAVEATS AND WARNINGS

- The estimates derived from the orphanhood method are conditional survivorship probabilities, that is to say probabilities of survival across an interval in adulthood conditional on being alive at the start of the interval. To obtain a complete life table, estimates of survivorship from birth to adulthood must be calculated using another source of data on child mortality.
- In a number of applications in East Africa and elsewhere, the orphanhood method has yielded results that indicate implausibly rapid declines in mortality and gross inconsistencies between the estimates from successive enquiries. This appears to be due to 'the adoption effect', that is under-reporting of orphanhood among those whose parents die when they are very young (Blacker 1984; Blacker and Gapere 1988; Hill 1984; Timæus 1986). Children who are orphaned at a young age tend to be reared by other relatives and are often enumerated as their own children. This means they are enumerated as having a living parent, and can give rise to very low mortality estimates. Misreporting appears to be particularly common when the mother dies. As the respondents get older, the chance that their foster, adoptive or step-parent

has died, as well as their biological parent, increases. This implies that the bias is most pronounced for young children, whose substitute parent is very likely to be alive. Procedures for estimating adult mortality from synthetic cohort data on orphanhood reported by young adults were developed specifically to address this problem, but cannot completely eliminate the bias in the results in populations where orphanhood is severely underreported.

- Although estimates can be made using data on respondents aged in their forties, the parents of many of these respondents are elderly and have very high mortality. This means that the precision with which one can estimate mortality from parental survival data is inherently much lower than it is for younger respondents.
- Like all methods that involve the analysis of change between two independent inquiries, the synthetic cohort approach to the analysis of orphanhood data is vulnerable to bias resulting from differences in data quality between the two inquiries. If respondents were more likely to report dead parents as living in one of the inquiries than the other, the resulting bias will be magnified in the synthetic cohort estimates. Mortality will be overestimated if too few orphans were reported in the earlier inquiry and underestimated if too few orphans were reported in the later inquiry. In addition, estimates based on the change in the proportion of parents that are alive in between two surveys have larger sampling errors than the two sets of proportions from which they are calculated.

## APPLICATION OF METHOD

### *Step 1: Calculate the mean ages of childbearing of women and men*

To apply the orphanhood method, one requires an estimate of the average age at which the parents had children in order to control for variation in the age range over which they have been exposed to the risk of dying. Women's mean age of childbearing is usually calculated from census or survey data on births in the last year by five-year age group at interview of the women giving birth. The measure is simply the average age of women giving birth calculated without adjusting for the age structure of the population using the following formula:

$$\bar{M}^f = \frac{\sum_{x=15}^{45} {}_5B_x\,(x+2)}{\sum_{x=15}^{45} {}_5B_x}.$$

In this equation, $(x+2)$ represents the mid-point of the age group of women with a half-year downward shift to allow for the fact that women giving birth in the year before interview did so 6 months ago, on average, and were 6 months younger at that time. This calculation can be done in the workbook (see website). If the data used to calculate $\bar{M}^f$ are tabulated by women's age at giving birth, the mid-point of each age group would become $x+25$.

There is no need to adjust the births data for reference-period errors before calculating $\bar{M}^f$. Moreover, the mortality estimates are not very sensitive to bias in this indicator. However, if evidence exists that the age pattern of births has been distorted severely by women exaggerating their ages, the number of births by age could be recomputed from an adjusted age distribution and adjusted fertility distribution before calculating $\bar{M}^f$.

In principle, the mean age of motherhood should refer to the time at which the respondents were born, which may be any time between 5 and 45 years before the collection of the orphanhood data. An estimate based on fertility data collected in the first of the pair of inquiries that asked about orphanhood should be adequate in populations which at that time had yet to experience substantial fertility decline. If fertility is believed to have fallen and earlier census or survey data exist, $\bar{M}^f$ could also be calculated from the earlier data to determine if it has changed. If it has, then the best way of deciding on final values of $\bar{M}^f$ for the estimation of adult mortality will depend on what data are available and the pattern of change in fertility. One option might be to calculate $\bar{M}^f$ from data collected at about the time that fertility began to fall and use that value for age groups of respondents born then or earlier and to interpolate linearly between that value and the current one to estimate $\bar{M}^f$ for younger age groups of respondents.

The mean age at which men have children is usually estimated by adding an index of the difference in the ages of men and women bearing children to the mean age of childbearing of women:

$$\bar{M}^m = \bar{M}^f + d.$$

One estimate of this difference that can be readily calculated from census data is the difference between the median ages of currently married men and currently married women. It is more appropriate than the difference between the singulate mean ages at marriage of men and women in populations in which marital dissolution or polygynous marriage is

common. The median is used rather than the mean so that differential age exaggeration by older respondents, who are probably no longer bearing children anyway, does not distort the estimate.

This approach to the estimation of the mean age of men at the birth of their children assumes that the ages of the fathers of children born to unmarried women are the same, on average, as the ages of the fathers of children born to married women. They may not be and this could introduce a significant bias into the estimate of $\bar{M}^m$ in populations in which childbearing outside marriage is common. While it is difficult to think of a solution to this problem, fortunately the mortality estimates are not very sensitive to errors in the estimate of $\bar{M}^m$.

***Step 2: Calculate the synthetic cohort measures of orphanhood in adulthood***

The workbook (see website) contains separate sheets for the calculation of these proportions for adult women and for adult men. Either the number of respondents by five-year age group with living mothers and the number answering the question or the proportions with living mothers calculated from them should be entered into the maternal orphanhood sheet. Similarly, either the number of respondents by five-year age group with living fathers and the number answering the question or the proportions with living fathers calculated from them should be entered into the paternal orphanhood sheet. For both the mothers and the fathers, the more recent set of results should be entered in the upper panel of the sheet and the more distant set in the lower panel. While the data can be for female respondents, male respondents, or respondents of both sexes, the two sets of results should usually be tabulated on the same basis.

The spreadsheet calculates the synthetic cohort measures of orphanhood by means of a 'variable *r*' method. The average proportions of respondents with living parents for the period between the two inquiries are multiplied by the exponential of the growth rates in those proportions during the period cumulated from age 20. This 'removes' the effect of population growth, producing stationary proportions relative to the proportion with living parents at the base age of 20. These stationary proportions reflect the rate at which adults are being orphaned during the period between the two inquiries.

The average proportion of respondents with living mothers (or fathers) in an age group over the period between the two inquiries is calculated as

$$ {}_5S_x(\bar{t}) = \sqrt{{}_5S_x(t) \times {}_5S_x(t+h)} $$

where $t$ indicates the first inquiry, $t+h$ the second inquiry occurring $h$ years later, and $\bar{t}$ a measure applying to the intervening period. Having calculated these measures, the average proportion of the parents of individuals aged exactly 20 that are alive during the period can be estimated as

$$ S(20,\bar{t}) = \sqrt{{}_5S_{15}(\bar{t}) \times {}_5S_{20}(\bar{t})}. $$

The growth rates in the proportions of parents that are alive by age group of respondent between the first and second inquiry are calculated as

$$ {}_5r_x(\bar{t}) = \frac{\ln({}_5S_x(t+h)) - \ln({}_5S_x(t))}{h}. $$

Then the synthetic cohort proportions that have living parents among those who had a living parent at age 20 can be calculated as

$$ \frac{{}_5S_x(\tau)}{S(20,\bar{t})} = \frac{{}_5S_x(\bar{t})}{S(20,\bar{t})} \cdot e^{5\left(\sum_{a=20,5}^{x-5} {}_5r_a(\bar{t})\right) + 2.5\,{}_5r_x(\bar{t})} $$

where $\tau$ indicates adjusted synthetic cohort (i.e. period) measures for time $\bar{t}$.

***Step 3a: Calculate the conditional life table survivorship ratios for women***

The survivorship of women is estimated between a lower age of 45 and age $25+n$, where $n$ is the upper limit of each successive age group of respondents. The following regression equation and the coefficients shown in Table 26.1 are used:

$$ {}_{n-20}p_{45} = a(n) + b(n)\bar{M}^f + c(n)\,{}_5S_{n-5}(\tau)/S(20,\bar{t}). $$

*TABLE 26.1 Coefficients for the estimation of women's survivorship from the proportions of adult respondents with living mothers among those with living mothers at age 20*

| *n* | *a*(*n*) | *b*(*n*) | *c*(*n*) |
|---|---|---|---|
| 25 | –0.8623 | 0.00292 | 1.7861 |
| 30 | –0.3822 | 0.00679 | 1.2062 |
| 35 | –0.4355 | 0.01197 | 1.1310 |
| 40 | –0.5995 | 0.01847 | 1.1419 |
| 45 | –0.7984 | 0.02547 | 1.1866 |
| 50 | –0.9360 | 0.03039 | 1.2226 |

Source: Timæus (1991)

For example, when $n$ is 30, life table survivorship is estimated over the ten-year age interval from exact age 45 to exact age 55 using data on survival of mothers supplied by respondents aged 25–29 years.

***Step 3b: Calculate the conditional life table survivorship ratios for men***

Each estimate of the survivorship of men is produced using data on two adjacent five-year age groups, not a single age group. Men's survivorship is measured from age 55 to 35 + $n$, where $n$ is the midpoint of the pair of age groups, using the following regression equation and the coefficients shown in Table 26.2:

$$ {}_{n-20}p_{55} = a(n) + b(n)\bar{M}^{m} + c(n)\,{}_{5}S_{n-5}(\tau)/S(20,\bar{t}) + d(n)\,{}_{5}S_{n}(\tau)/S(20,\bar{t}). $$

For example, when $n$ is 40, life table survivorship is estimated over the 20-year age interval from exact age 55 to exact age 75 using the data on survival of fathers supplied by respondents in the two age groups 35–39 years and 40–44 years.

*TABLE 26.2 Coefficients for the estimation of men's survivorship from the proportions of adult respondents with living fathers among those with living fathers at age 20*

| $n$ | $a(n)$ | $b(n)$ | $c(n)$ | $d(n)$ |
|---|---|---|---|---|
| 25 | –0.0554 | 0.00757 | 0.0239 | 0.8080 |
| 30 | –0.7539 | 0.01558 | 0.6452 | 0.6498 |
| 35 | –1.0809 | 0.02273 | 0.9289 | 0.4807 |
| 40 | –1.1726 | 0.02647 | 0.9381 | 0.4372 |

Source: Timæus (1991)

***Step 4: Convert the survivorship ratios into estimates of the level of mortality***

The series of conditional survivorship ratios, ${}_{n}p_{b}$, obtained from different age groups of respondents all refer to the interval between the two surveys. They represent incomplete life tables with a base at age 45 for women and age 55 for men. The series will be to some extent erratic as a result of age reporting, sampling, and other errors. It can be smoothed by fitting a 2-parameter logit model life table to the ratios. The logits of the conditional survivorship ratios are calculated as

$$ Y_{x} = \frac{1}{2}\ln\left(\frac{1 - {}_{n}p_{b}}{{}_{n}p_{b}}\right). $$

The equivalent logits of the standard life table are

$$ Y_{x}^{s} = \frac{1}{2}\ln\left(\frac{l_{b} - l_{b+n}}{l_{b+n}}\right). $$

The $\alpha$ and $\beta$ parameters of the fitted model are the intercept and slope of a regression of the $Y_{x}$ values on $Y_{x}^{s}$. In principle, the estimates for older age groups are less vulnerable to sampling error than those on younger age groups as they are based on more parental deaths. However, these estimates can indicate lower mortality than the estimates for younger age groups, perhaps because the respondents are exaggerating their ages. Thus, one should exclude any estimates at either end of the series that are out of line with the others from the range of ages used to estimate $\alpha$ and $\beta$.

Once one has calculated $\alpha$ and $\beta$, smoothed estimates of conditional survivorship can be calculated as

$$ {}_{n}\hat{p}_{b} = \frac{1}{1 + \exp\left(2\left(\alpha + \beta Y_{x}^{s}\right)\right)}. $$

The smoothed estimates of conditional survivorship refer to a clearly defined period of time and depend only to a limited extent on assumptions made during the estimation process about the age pattern of mortality. They will not be distorted greatly in populations with unusual age patterns of mortality such as those experiencing generalized HIV epidemics. Thus, if possible, they should be used as they are in further analyses. Nevertheless, it is often necessary to convert the synthetic cohort estimates into a common index of mortality in order to compare those for men and women directly or to compare both series with estimates of mortality from other sources. This can be done by fitting a 1-parameter model life table to each conditional survivorship ratio and obtaining the desired index from the fitted model.

A wide range of indices have been used for this purpose, including the level parameters of various systems of model life tables, survivorship ratios, life expectancy at various ages between 5 and 30, and temporary life expectancy between ages 25 and 70, ${}_{45}e_{25}$. Using the parameters of the models has the advantage of emphasizing that the full life table is being estimated by fitting a model, rather than measured directly. The measures of life expectancy summarize survivorship across adulthood as a whole, while using survivorship ratios or temporary life expectancies avoids extrapolation into old age from measures for younger adults. Increasingly, in recent years, the estimates have been presented in terms of the probability of dying between exact ages 15 and 60, ${}_{45}q_{15}$,

as this measure has found favour with several international agencies as a summary indicator of the mortality of young and middle-aged adults.

In the applications of the orphanhood method presented here the survivorship ratios are converted into the $\alpha$ parameter of a 1-parameter system of logit model life tables, and then into estimates of the conditional probability of dying across a wider range of ages. (Note that, even if the same standard is used and $\beta$ is 1, the $\alpha$ parameter of a fitted model based at age 0 will not be the same as $\alpha$ in models that have been fitted to measures of conditional survivorship from age 45 or 55.)

The spreadsheet (see website) calculates conditional survivorship between exact ages 30 and 60, ${}_{30}q_{30}$, exact ages 15 and 60, ${}_{45}q_{15}$, or exact ages 50 and 70, ${}_{20}q_{50}$. The first two indices are useful for comparing the synthetic cohort estimates with those from the basic orphanhood method and other adult mortality measures respectively; the third is most useful for comparing the estimates made from orphanhood in adulthood for men and women or for assessing the internal consistency of a series of such estimates without extrapolating from survivorship in middle age to younger adult ages. The parameters of the 1-parameter models are calculated from the estimates of ${}_{n-20}p_b$ as

$$\alpha = -\frac{1}{2}\ln\left(1+\frac{\frac{{}_{n-20}p_b}{l^s_{b+n-20}}-\frac{1}{l^s_b}}{1-{}_{n-20}p_b}\right)$$

where the estimates of ${}_{n-20}p_b$ come from Step 2, with $b=45$ for the estimates of the survivorship of women and $b=55$ for those of the survivorship of men, and the $l^s_x$ values come from a standard life table. Thus, one obtains a series of estimates of $\alpha$ corresponding to the measures of conditional survivorship made from data on the different age groups of respondents. Higher values of $\alpha$ correspond to higher mortality. Then, for each$\alpha$, summary measures such as ${}_{20}q_{50}$, ${}_{30}q_{30}$ and ${}_{45}q_{15}$ can be calculated as

$${}_nq_b = 1-\frac{1+e^{2\left(\alpha+Y^s(b)\right)}}{1+e^{2\left(\alpha+Y^s(b+n)\right)}}.$$

The spreadsheet (see website) can calculate these measures using either a standard from the General set of United Nations model life tables or one from any of the four families of Princeton model life tables. The standard life table should be chosen to have an age pattern of mortality within adulthood that resembles that of the population being studied. Another life tables can be used as a standard instead if there is reason to believe that it resembles more closely the pattern of adult mortality in the population being studied. The most suitable life table may not be from the family of models that best captures the relationship between child and adult mortality. If nothing is known about the age pattern of mortality in adulthood, use of the United Nations General or Princeton West models is recommended.

As the estimates all refer to the same period, it makes sense to produce the final estimate of survivorship for the period between the two inquiries by averaging a contiguous set of estimates that excludes any outlying values made from data on the youngest and oldest respondents. Such outliers can be identified in a plot of the logits of the conditional survivorship ratios against a standard series. If there is a clear upward or downward trend in $\alpha$ across the age groups in the fitted 1-parameter models, the mortality standard to which the estimates are being fitted may be inappropriate. The analysis should probably either adopt another standard or modify the rate at which mortality increases with age in the selected one by adjusting its $\beta$ parameter.

***Step 5: Calculate the time location of the estimates***

Each survivorship ratio refers to the period between the dates to which the two sets of orphanhood data refer. One may wish to ascribe them to an exact date within this period, so that they can be plotted and compared with other estimates of adult mortality. If one assumes a constant rate of change in mortality, they can be thought of as referring to the geometric average of the dates of the two inquiries. The date of each inquiry can be calculated as the average of the dates on which the interviews took place or taken as the mid-point of the period of fieldwork if exact dates of interview are not available.

## WORKED EXAMPLE

This example, implemented in the workbook (see website), uses data on the survival of mothers and fathers collected in the 1989 and 1999 Censuses of Kenya.

***Step 1: Calculate the mean ages of childbearing of women and men***

For women the mean age of childbearing is a straightforward average of the ages of women giving birth and can either be calculated as such from individual-level data or estimated

approximately from a tabulation of births by five-year age group of mother. For this application it has been calculated using data from the first of the pair of censuses (see Table 26.3), although in Kenya one could also do so using data from previous censuses to check whether ages at childbearing have changed over time:

$$\bar{M}^f = \frac{16{,}585{,}300}{620{,}000} = 26.75.$$

*TABLE 26.3 Calculation of the mean age at childbearing, Kenya, 1989 Census*

| Age group | Births in the last year $B(i)$ | Mid-point age $N$ | $B(i)*N$ |
|---|---|---|---|
| 15–19 | 73,600 | 17 | 1,251,200 |
| 20–24 | 193,400 | 22 | 4,254,800 |
| 25–29 | 170,220 | 27 | 4,595,940 |
| 30–34 | 95,180 | 32 | 3,045,760 |
| 35–39 | 56,340 | 37 | 2,084,580 |
| 40–44 | 23,240 | 42 | 976,080 |
| 45–49 | 8,020 | 47 | 376,940 |
| **TOTALS** | **620,000** | | **16,585,300** |

The mean age of childbearing of men is calculated by adding the difference between the median ages of currently married men and women to the mean age of childbearing of women. It can be seen from Table 26.4 that the median age of currently married men falls between the mid-point of the age group 30–34 and the mid-point of the age group 35–39. By linear interpolation:

$$\text{Median for men} = 32.5 + 5 \times \frac{0.5 - 0.3886}{0.5272 - 0.3886} = 36.52$$

and

$$\text{Median for women} = 27.5 + 5 \times \frac{0.5 - 0.4234}{0.5597 - 0.4234} = 30.31.$$

Then the estimated mean age of childbearing of men is

$$M^m = M^f + d = 26.75 + (36.52 - 30.31) = 32.96.$$

***Step 2: Calculate the synthetic cohort measures of orphanhood in adulthood***

Proportions of Kenyans with living mothers averaged across the period between the 1989 and 1999 Censuses of Kenya are shown in the fourth column of Table 26.5. They are the

*TABLE 26.4 Ages of currently married men and women, Kenya, 1989 Census*

| Age group | Married men | Married women | Cumulative proportion of men | Cumulative proportion of women |
|---|---|---|---|---|
| 10–14 | 2,800 | 6,680 | 0.0010 | 0.0019 |
| 15–19 | 18,040 | 212,060 | 0.0071 | 0.0612 |
| 20–24 | 173,840 | 623,040 | 0.0664 | 0.2356 |
| 25–29 | 464,720 | 670,760 | 0.2250 | 0.4234 |
| 30–34 | 479,460 | 487,180 | 0.3886 | 0.5597 |
| 35–39 | 406,000 | 387,000 | 0.5272 | 0.6681 |
| 40–44 | 330,140 | 305,500 | 0.6398 | 0.7536 |
| 45–49 | 250,540 | 243,120 | 0.7253 | 0.8216 |
| 50–54 | 212,820 | 189,240 | 0.7979 | 0.8746 |
| 55–59 | 161,760 | 137,120 | 0.8531 | 0.9130 |
| 60–64 | 135,060 | 113,860 | 0.8992 | 0.9449 |
| 65–69 | 101,860 | 75,540 | 0.9340 | 0.9660 |
| 70–74 | 72,080 | 49,980 | 0.9586 | 0.9800 |
| 75–79 | 56,240 | 30,100 | 0.9778 | 0.9884 |
| 80+ | 65,120 | 41,380 | 1.0000 | 1.0000 |
| **TOTAL** | **2,930,480** | **3,572,560** | | |

geometric averages of the proportions reported in the two censuses shown in the second and third columns of the table. For example, in the age group 25–29,

$${}_5S_{25}(\bar{t}) = \sqrt{0.8839 \times 0.8771} = 0.8805.$$

The proportion with living mothers at exact age 20 is calculated from these estimates for age groups 15–19 and 20–24:

$$S(20, \bar{t}) = \sqrt{0.9446 \times 0.9156} = 0.9300.$$

Census day in 1999 was 24th August, while in 1989 it was 25th October. Thus, the growth rate over the decade in the proportion of Kenyans with living mothers for the same age group is

$${}_5r_{25}(\bar{t}) = \frac{\ln(0.8771) - \ln(0.8839)}{9.83} = -0.00078.$$

For the first age group, the growth rate cumulated from age 20 to 22.5 is simply

$$2.5\,{}_5r_{20}(\bar{t}) = -0.00170 \times 2.5 = -0.00425.$$

For the second age group, the growth rate cumulated to age 20 to 27.5 is

$$5\sum_{a=20,5}^{20} {}_5r_a(\bar{t}) + 2.5\,{}_5r_{25}(\bar{t})$$
$$= 5\times -0.00170 + 2.5\times -0.00078 = -0.01045$$

while for the third age group it is

$$5\sum_{a=20,5}^{25} {}_5r_a(\bar{t}) + 2.5\,{}_5r_{30}(\bar{t})$$
$$= 5\times(-0.00170 - 0.00078) + 2.5\times 0.00018 = -0.01195$$

and so on.

The synthetic cohort proportions that have living mothers among those who had a living mother at age 20 in the seventh column of Table 26.5 are calculated from the averaged proportions and growth rates in the fourth and fifth columns. For example, for the 25–29 age group,

$$\frac{{}_5S_x(\tau)}{S(20,\tau)} = \frac{0.8805}{0.9300}\cdot e^{-0.01045} = 0.9369.$$

The calculations made in this step for the data on paternal orphanhood are identical and are shown in Table 26.6.

***Step 3a: Calculate the conditional life table survivorship ratios of women***

These survivorship ratios are shown in the eighth column of Table 26.5 and are calculated from the proportions in the seventh column using the regression coefficients shown in Table 26.1 and the estimate of $\bar{M}^f$ of 26.75 from Step 1. For example, for respondents aged 25–29,

$${}_{10}p_{45} = -0.3822 + 0.00679 \times 26.75 + 1.2062\times 0.9369 = 0.9295.$$

Note that each life table measure is similar in value to the proportion from which it was calculated.

***Step 3b: Calculate the conditional life table survivorship ratios of men***

These survivorship ratios are shown in the eighth column of Table 26.6 and are calculated from the proportions in the seventh column using the regression coefficients shown in Table 26.2 and the estimate of $\bar{M}^m$ of 32.96 from Step 1. For example, for the final estimate in Table 26.6:

$${}_{20}p_{55} = -1.1726 + 0.02647\times 32.96 + 0.9381 \times 0.6156 + 0.4372\times 0.4772 = 0.4860.$$

***Step 4: Convert the survivorship ratios into estimates of the level of mortality***

To smooth the series of estimates of conditional survivorship by fitting a 2-parameter logit model life table to them, one first calculates the logits of the ratios. For example, the estimate of ${}_{10}p_{45}$ for women made from data on respondents aged 25–29 is

$$Y_x = \frac{1}{2}\ln\left(\frac{1-0.9295}{0.9295}\right) = -1.290.$$

*TABLE 26.5 Estimation of women's survivorship in the interval between two inquiries, and corresponding estimates of $\alpha$ and ${}_{30}q_{30}$, from maternal orphanhood in adulthood, Kenya, 1989–1999*

| Age group | Proportion alive 1989 ${}_5S_{n-5}(t)$ | Proportion alive 1999 ${}_5S_{n-5}(t+h)$ | Proportion alive ${}_5S_{n-5}(\bar{t})$ | Growth rate | Age $n$ | Proportion alive ${}_5S_{n-5}(\tau)$ | Estimated $\frac{l(25+n)}{l(45)}$ | Smoothed $\frac{l(25+n)}{l(45)}$ | Probability of dying ${}_{30}q_{30}$ |
|---|---|---|---|---|---|---|---|---|---|
| 15–19 | 0.9557 | 0.9336 | 0.9446 | | | | | | |
| 20–24 | 0.9233 | 0.9080 | 0.9156 | –0.00170 | 25 | 0.9804 | 0.9669 | 0.9667 | 0.192 |
| 25–29 | 0.8839 | 0.8771 | 0.8805 | –0.00078 | 30 | 0.9369 | 0.9295 | 0.9291 | 0.172 |
| 30–34 | 0.8229 | 0.8244 | 0.8236 | 0.00018 | 35 | 0.8751 | 0.8745 | 0.8804 | 0.167 |
| 35–39 | 0.7553 | 0.7691 | 0.7622 | 0.00184 | 40 | 0.8139 | 0.8240 | 0.8145 | 0.140 |
| 40–44 | 0.6258 | 0.6685 | 0.6468 | 0.00671 | 45 | 0.7057 | 0.7203 | 0.7244 | 0.140 |
| 45–49 | 0.5335 | 0.5653 | 0.5492 | 0.00589 | 50 | 0.6184 | 0.6329 | 0.6037 | 0.113 |

The equivalent value for the UN General model life table with $e_0 = 60$ is

$$Y_x^s = \frac{1}{2}\ln\left(\frac{0.7887 - 0.7103}{0.7103}\right) = -1.102\,.$$

Regressing the logits of the observed estimates of conditional survivorship on the standard logits, excluding the final point (based on respondents aged 45–49) which dips below the line, gives parameter estimates of $\alpha = -0.3398$ and $\beta = 0.8597$. The fact that the estimate of $\beta$ is less than 1 indicates that the mortality of women in Kenya increases less steeply than in the standard across the age range 45 to 75 years.

Having obtained $\alpha$ and $\beta$, the smoothed conditional survivorship ratio for the second age group, for example, can be calculated as

$$\begin{aligned} {}_{10}\hat{p}_{45} &= \frac{1}{1+\exp\left(2\left(-0.3398 + 0.8597 \times -1.1016\right)\right)} \\ &= 0.9291. \end{aligned}$$

The full series of fitted survivorship ratios is shown in the ninth column of Table 26.5 for women and of Table 26.6 for men. Taking the estimates of ${}_5p_{45}$ and ${}_{25}p_{45}$ from Table 26.5, the conditional probability of dying between exact ages 50 and 70 in the fitted 2-parameter model is $1 - 0.7244/0.9667 = 0.251$.

Estimates of ${}_{30}q_{30}$, the probability of dying between exact ages 30 and 60, calculated by fitting 1-parameter models to the estimated survivorship ratios, are shown in rightmost columns of Table 26.5 and Table 26.6. For example, $\alpha$ is calculated from the estimate of ${}_{10}p_{45}$ for women as

$$\alpha = -\frac{1}{2}\ln\left(1 + \frac{\dfrac{0.9295}{0.7103} - \dfrac{1}{0.7887}}{1 - 0.9295}\right) = -0.228\,.$$

Having calculated $\alpha$, then the corresponding measure of ${}_{30}q_{30}$ is

$${}_{30}q_{30} = 1 - \frac{1 + e^{2(-0.228 - 0.8655)}}{1 + e^{2(-0.228 - 0.3064)}} = 0.172\,.$$

***Step 5: Calculate the time location of the estimates***

The synthetic cohort estimates are measures of adult mortality during the period between the two inquiries. Their time location can be ascribed to the geometric average of the dates of field work of the two inquiries. Thus, in this application to the 1989 and 1999 Censuses of Kenya:

$$T = \sqrt{1989.81 \times 1999.64} = 1994.72.$$

## DIAGNOSTICS, ANALYSIS AND INTERPRETATION

### Checks and validation

The number of respondents who stated that they did not know whether their mother or father is alive or who did not answer the questions at all should be checked before they are dropped from the analysis. If many of the respondents failed to respond to these questions, the data supplied by those respondents who did answer them may not be representative of the population as a whole. Moreover, a high level of non-response may indicate that either the field staff or the respondents experienced difficulty with the questions. Thus, even when answers were supplied they may be rather

*TABLE 26.6 Estimation of men's survivorship in the interval between two inquiries, and corresponding estimates of $\alpha$ and ${}_{30}q_{30}$, from paternal orphanhood in adulthood, Kenya, 1989–1999*

| Age group | Proportion alive 1989 ${}_5S_{n-5}(t)$ | Proportion alive 1999 ${}_5S_{n-5}(t+h)$ | Proportion alive ${}_5S_{n-5}(\bar{t})$ | Growth rate | Age $n$ | Proportion alive ${}_5S_{n-5}(\tau)$ | Estimated $\frac{l(35+n)}{l(55)}$ | Smoothed $\frac{l(35+n)}{l(55)}$ | Probability of dying $({}_{30}q_{30})$ |
|---|---|---|---|---|---|---|---|---|---|
| 15–19 | 0.8670 | 0.8368 | 0.8518 | | | | | | |
| 20–24 | 0.7971 | 0.7730 | 0.7849 | –0.00312 | 25 | 0.9525 | 0.9052 | 0.9045 | 0.259 |
| 25–29 | 0.7136 | 0.7055 | 0.7096 | –0.00117 | 30 | 0.8519 | 0.7816 | 0.7849 | 0.257 |
| 30–34 | 0.6071 | 0.6074 | 0.6073 | 0.00004 | 35 | 0.7270 | 0.6395 | 0.6370 | 0.244 |
| 35–39 | 0.4972 | 0.5198 | 0.5084 | 0.00453 | 40 | 0.6156 | 0.4860 | 0.4678 | 0.225 |
| 40–44 | 0.3729 | 0.3953 | 0.3839 | 0.00592 | | 0.4772 | | | |

unreliable. If a high level of item non-response exists, it can be illuminating to determine whether it is concentrated among a minority of field staff or certain type of respondent.

If information about the survival of mothers and fathers has been collected from both male and female *respondents* in a census or a large-scale survey with small sampling errors, it is possible to tabulate the proportions of mothers and fathers alive separately for respondents of each sex in order to compare the consistency of their reports. While consistency of reporting does not guarantee accuracy, statistically significant differences between the proportions obtained from male or female respondents imply that at least one sex, and possibly both of them, are answering the questions inaccurately.

It is fairly common to find that women report lower proportions of living parents than men. Some analysts believe that this is because women stay in closer contact with their parents than men and that some men are stating that their parents are alive because they do not know that they have in fact died. If correct, this would imply that the data supplied by women are more accurate. However, no strong evidence exists to support this interpretation and other errors, such as differential age misreporting by male and female respondents, may also produce inconsistencies between the proportions reported by men and women. Because synthetic cohort estimates are calculated, in essence, from the first differences between two sets of proportions, they have larger sampling errors than other orphanhood-based estimates of mortality. Thus, especially if the data come from surveys of a few thousand households, it is advisable to base the final estimates of adult mortality on the combined responses both sexes unless clear evidence exists that one should focus on the data supplied by female respondents.

**Interpretation**

The results of the example analysis of the orphanhood data from the 1989 and 1999 Censuses of Kenya are portrayed graphically in Figure 26.1. According to the 1989 Census data on orphanhood, adult mortality in Kenya was declining slowly in the late 1970s and early 1980s. The level of mortality was fairly low and a large differential existed between the mortality of men and women. By contrast, the 1999 Census data suggest that mortality rose steadily for both men and women to a considerably higher level between the late-1980s and mid-1990s.

One reassuring feature of these results is that the mortality estimates for 1985–1986 from the two censuses are consistent. Those from the 1989 Census (the most recent points on the dotted lines) are based on reports about the survival of the parents of respondents who are still children. Those for only slightly later made from the 1989 Census (the earliest points on the solid lines) are made from the reports of respondents who were aged in their thirties in 1999. While such consistency between estimates made from the reports of respondents of different ages in different inquiries does not guarantee their accuracy, it is suggestive of it and rules out the presence of certain (but not all possible) errors, including bias resulting from the adoption effect. This effect is most severe for estimates made from data on children because, the younger a child is when its parent dies, the more likely it is that a question about whether the parent is alive will be answered with reference to an adoptive, foster, or step parent who has reared them. As the respondents get older, these misreported cases become proportionately less important compared with the rapidly increasing number of parental deaths that occur as both the respondents and their parents get older. Thus, if the adoption effect was a problem in Kenya one would expect the 1989 Census estimates of adult mortality in the mid-1980s to be lower than the 1999 Census estimates for the same years.

If the mortality estimates obtained from younger respondents by the basic orphanhood method were biased downward, one would expect the synthetic cohort estimates of mortality derived from orphanhood data supplied by young adults to be higher than those from the basic method for the same dates. In Kenya, they are not – they are lower.

This pattern of synthetic data on young adults yielding lower estimates than lifetime data on children is unusual. It probably reflects the growing importance of AIDS mortality in Kenya during the course of the 1990s. The synthetic cohort estimates are based mainly on the experience of parents aged 50 years of more, who have not been affected greatly by the AIDS epidemic. Thus, using standard model life tables to determine ${}_{30}q_{30}$ from these data produces underestimates because mortality at ages 30 to 50 in the fitted models is lower than in Kenya.

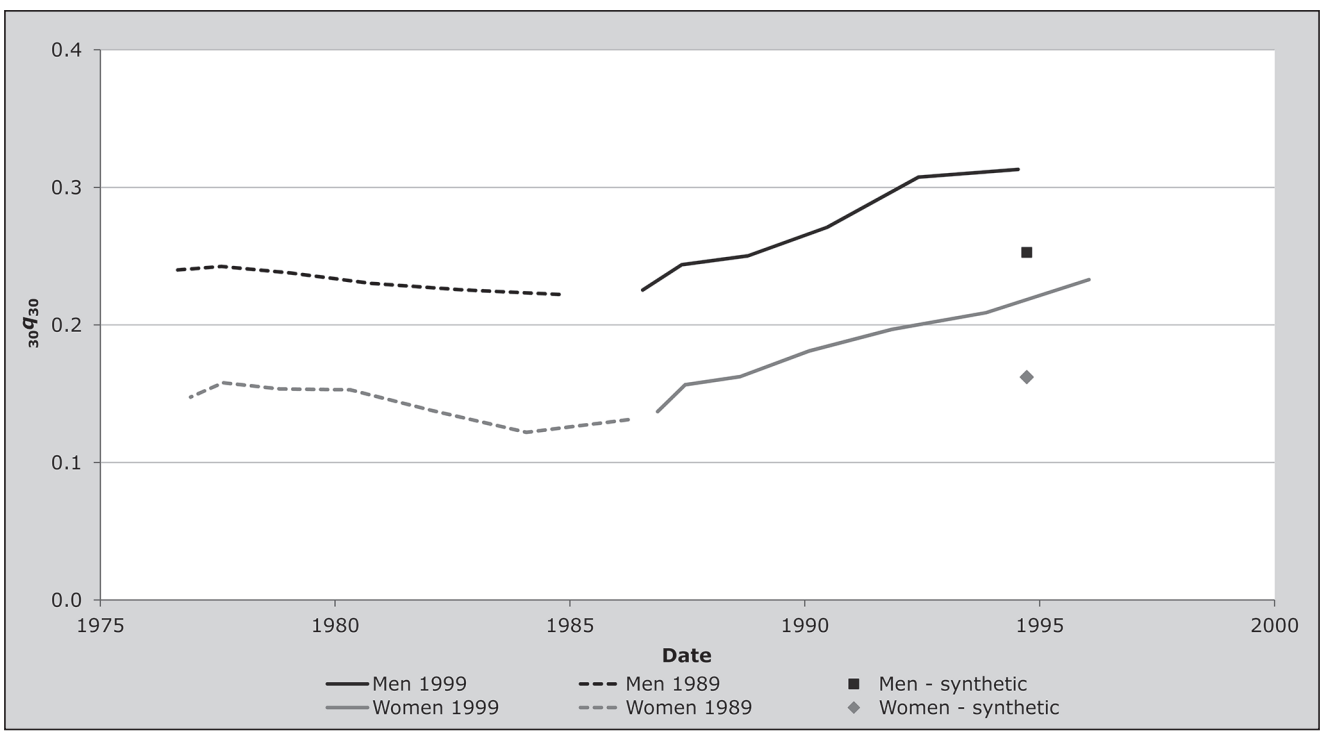
In contrast, the parents of the young respondents, whose reports are the basis for the most up-to-date estimates obtained from the basic orphanhood method, are largely in their 30s and 40s. AIDS mortality peaks in this age range. Thus, using standard model life tables to determine ${}_{30}q_{30}$ from these data produces overestimates because mortality at ages 50 to 60 in the fitted models is higher than in Kenya.

The synthetic cohort estimates provide support for this interpretation. In this application, the conditional probability of dying between ages 30 and 60 declines as the age group of the respondents increases for both men and women (see Table 26.5 and Table 26.6). This suggests that mortality is relatively high at younger ages within the age range 45 to 75 years and relatively low in late middle age in Kenya compared with United Nations' General family of model life tables. This is also suggestive of high AIDS mortality among younger adults. The actual value of ${}_{30}q_{30}$ in the mid-1990s probably falls somewhere between the estimates produced by the two variants of the orphanhood method. Adopting a standard age pattern of mortality that increases more slowly with age, by setting the $\beta$ parameter of the standard life table to 0.7 and recalculating $\alpha$, produces an internally more consistent set of indices for both of the sexes. It also reduces the inconsistencies between the synthetic cohort estimates and the most up-to-date estimates obtained from the 1999 Census data using the basic orphanhood method. Thus, the probability of dying between ages 30 and 60 in Kenya in the mid-1990s, conditional on surviving to age 30, was probably about 20 per cent for women and 30 per cent for men.

Figure 26.2 presents a second application of the synthetic cohort approach to analysing two sets of orphanhood data collected in successive inquiries. It analyses data from the 1986 and 1999 Censuses of the Solomon Islands. In this application, the statistic plotted is the probability of dying between ages 15 and 60 conditional on being alive at age 15 (${}_{45}q_{15}$). One immediately obvious contrast between these series of estimates and those for Kenya and the Arab countries for which results are presented in the discussion of the basic orphanhood method is that they suggest that gender inequalities in adult mortality in the Solomon Islands are small.

This application of the estimation method provides clear evidence of problems with the orphanhood data collected in the 1986 Census. First, the estimates suggest that mortality was declining very rapidly but the most recent of the earlier series of estimates, based on data on children, indicate substantially lower mortality than the estimates for a couple

***FIGURE 26.1*** *Trends in the conditional probability of dying between exact ages 30 and 60 estimated from orphanhood, Kenya, 1989 and 1999 Censuses*



of years later that were made using data collected from older respondents in the 1989 Census. Inconsistencies of this sort usually indicate that the more recent estimates from the first of the censuses are too low because orphanhood of children is being underreported due to the adoption effect. Because the tendency to underreport in this way may be an enduring feature of the culture of a population, such inconsistencies also cast doubt on the most recent estimates made from data collected in the later census. The second problem with the estimates from the 1986 Census of the Solomon Islands is that they suggest that women have higher adult mortality than men. This is very uncommon.

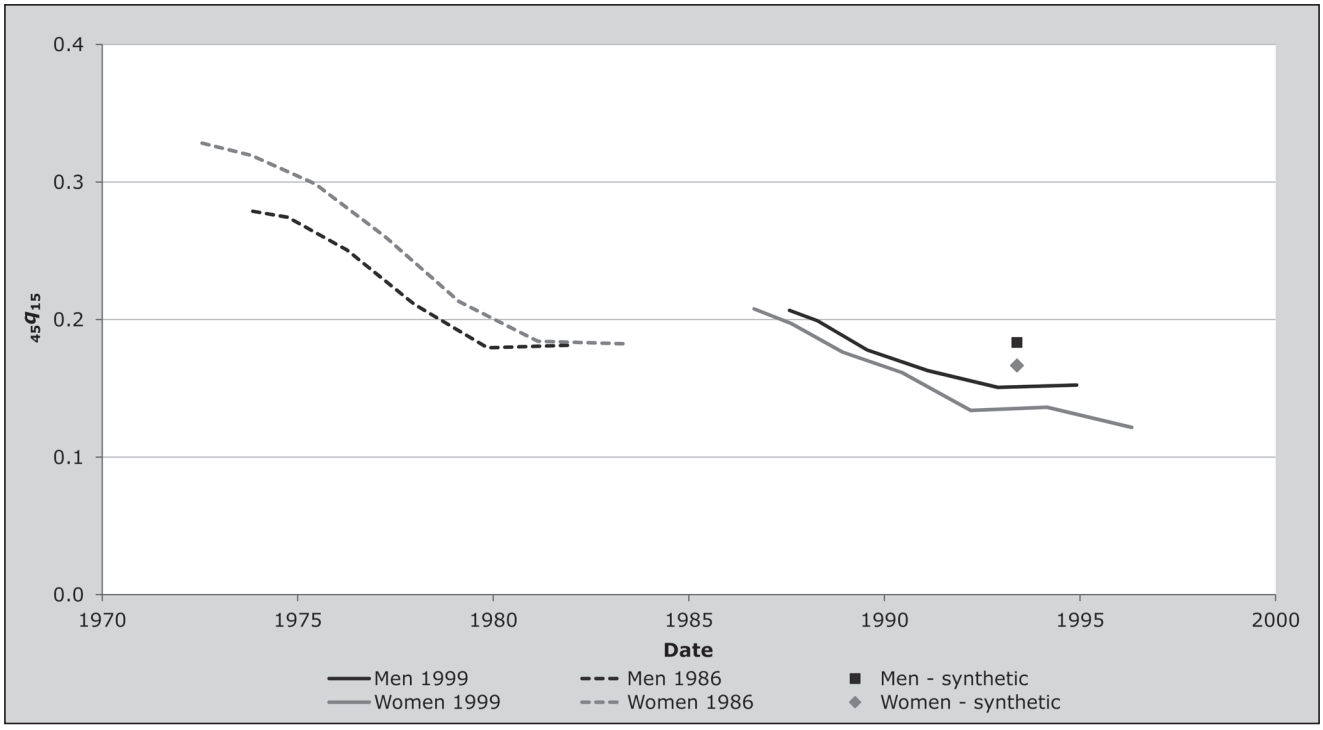
The estimates made for the early 1990s from synthetic cohort data on orphanhood in early adulthood support the suggestion that the lifetime estimates calculated from orphanhood data on children collected in 1999 are also too low. Because they are based exclusively on the reports of adults the synthetic cohort estimates are probably the most reliable estimates presented in Figure 26.2. Thus, it can be tentatively concluded that the probability of dying between ages 15 and 60 in the Solomon Islands fell from about 30 per cent to about 17.5 per cent in the two decades starting in the early 1970s.

## DETAILED DESCRIPTION OF METHOD

### Introduction

Simple, robust methods for estimating mortality from cohort data on orphanhood collected in a single inquiry were first published in Brass and Hill (1973). Zlotnik and Hill (1981) were the first to point out that, once the question 'Is your mother alive?' or 'Is your father alive?' has been asked in two successive inquiries in the same population, it becomes possible to calculate synthetic cohort measures of parental survival from the two sets of answers that reflect adult mortality in the intervening period. Because these estimates are made from changes in parental survival between the two inquiries, they are vulnerable to differential reporting and sampling errors. The time reference of the measures is usually more recent, however, than that of any of the estimates based on respondents' lifetime experience.

*FIGURE 26.2 Trends in the probability of surviving from exact age 15 to 60 estimated from orphanhood, Solomon Islands, 1986 and 1999 Censuses*



Synthetic cohort data also have another advantage: if deaths occurring during the period between the two inquiries are reported fully, omission of more distant deaths will have no impact on the results. Thus, synthetic cohort data on the survival of parents are potentially less vulnerable than lifetime data to the so-called 'adoption effect', that is underreporting of orphanhood by respondents whose parents died when the respondents were still young children. This is important because the adoption effect is the major bias affecting the orphanhood method, explaining the implausible results and inconsistencies between successive surveys found in a number of applications of this method.

The most straightforward way of purging the reports of this bias is to analyse only the synthetic cohort data on adults (Timæus 1991). One can do this by constructing a synthetic cohort, based at 20 years of age, from data on parental survival at two dates. This cohort indicates the proportion of the adult population whose mothers or fathers would remain alive, at current levels of mortality, among those who, at exact age 20, had a living mother or father. Such a synthetic cohort can be constructed solely from the relatively reliable data supplied by young adults.

Timæus (1991) proposes basing the cohort at 20 years for two reasons. First, this choice minimizes the possibility of underestimating orphanhood at the base age and consequently overestimating subsequent orphanhood and adult mortality. Second, because information on two age groups is needed to estimate parental survival at the exact age dividing the groups, this approach makes it possible to apply the method to data collected in surveys in which only women aged 15 to 49 are asked about orphanhood.

The generalization to all populations of the relationships between age structure, increase, and mortality stated by stable population theory (Preston and Coale 1982) provides a convenient way of constructing such synthetic cohorts. Stationary synthetic cohort measures of parental survival can be obtained by adjusting the reported proportions with living parents using the age-specific growth rates in these proportions to remove the impact of past trends in mortality. When the data come from two inquiries, adjustment using age-specific rates of increase in parental survival has the advantage over methods based on chaining cohort changes of being easy to apply to an interval between the inquiries of other than five and ten years.

**Mathematical exposition**

Preston and Coale (1982) show that in any closed population defined by age,

$$N(0,t) = N(a,t) \cdot e^{\int_0^a \mu(z,t)dz} \cdot e^{\int_0^a r(z,t)dz} \qquad (1)$$

where $N(a,t)$ is the number of individuals aged $a$ at time $t$ and $\mu(z,t)$ and $r(z,t)$ are the force of mortality and rate of growth respectively at age $z$ and time $t$. Attrition of the population with living mothers or living fathers, denoted $NO$, can be decomposed into the mortality of the parents and the mortality of the population itself (Preston and Chen 1984; Timæus 1986):

$$NO(0,t) = NO(a,t) \cdot e^{\int_0^a \mu(z,t)dz} \cdot e^{\int_0^a \pi(z,t)dz} \cdot e^{\int_0^a r^{NO}(z,t)dz}$$

where $\pi(z,t)$ represents the instantaneous rate of orphanhood, and $r^{NO}(z,t)$ the rate of growth of the population with living parents, at age $z$ and time $t$. Assuming that orphans and the rest of the population have identical mortality and using the fact that $N(0,t) \equiv NO(0,t)$, division of the expression for non-orphans by that for the total population produces

$$e^{-\int_0^a \pi(z,t)dz} = \frac{NO(a,t)}{N(a,t)} \cdot e^{\int_0^a r^{NO}(z,t) - r(z,t)dz} . \qquad (2)$$

Taken as a whole, the left-hand term represents the stationary probability of an individual aged $a$ having a living mother or father, denoted $S(a,\tau)$, whereas $NO(a,t)/N(a,t)$ equals the equivalent unadjusted proportion, $S(a,t)$. With survey data, it is more convenient to work with the rate of increase in the proportion of the population with living parents, $r^s(z,t)$, than with its equivalent, $r^{NO}(z,t) - r(z,t)$, the difference between the rates of increase for the non-orphaned and the total populations.

The population above any given age can be treated as self-contained and the relationship between age structure, increase, and mortality stated in Equation 1 will continue to hold for such populations. Thus, using the notation already established,

$$\frac{NO(20,t)}{N(20,t)} e^{-\int_{20}^a \pi(z,t)dz} = \frac{NO(a,t)}{N(a,t)} \cdot e^{\int_{20}^a r^s(z,t)dz}$$

for $a > 20$. When both sides are divided by $S(20,t)$ this becomes

$$\frac{S(a,\tau)}{S(20,\tau)} = \frac{S(a,t)}{S(20,t)} \cdot e^{\int_{20}^{a} r^s(z,t)dz}.$$

In discrete form, for age groups $x$ to $x+5$,

$$\frac{{}_5S_x(\tau)}{S(20,\tau)} = \frac{{}_5S_x(t)}{S(20,t)} \cdot e^{5\left(\sum_{z=20,5}^{x-5} {}_5r_z(t)\right)+2.5\,{}_5r_x(t)}.$$

**Implementation of the method**

In order to simplify the estimation of life table measures of mortality from these proportions, Timæus (1991) developed regression models for both men's and women's mortality estimating the coefficients from data on parental survival in the same set of simulated populations used to estimate those for the basic orphanhood method (Timæus 1992).

The proportion of individuals aged $a$ that have living mothers, $S(a)$, can be calculated as the average of the probabilities of surviving among mothers who gave birth at each age $y$, weighting by the proportion of births that occur at $y$ (Brass and Hill 1973),

$$S(a) = \frac{\int_s^\omega e^{-ry} f(y) l(y+a) dy da}{\int_s^\omega e^{-ry} f(y) l(y) dy}$$

where integration is over all ages at child bearing $s$ to $\omega$. Dividing $S(a)$ by $S(20)$ for $a > 20$, the denominators cancel. Thus, the proportion of a five-year age group with living mothers among those who had a living mother at exact age 20 is

$$\frac{{}_5S_x}{S(20)} = \frac{\int_x^{x+5} e^{-ra} l(a) \int_s^\omega e^{-ry} f(y) l(y+a) dy da}{\int_x^{x+5} e^{-ra} l(a) da \int_s^\omega e^{-ry} f(y) l(y+20) dy} \quad (3)$$

for $x \geq 20$. The equivalent proportion in each age group with living fathers is

$$\frac{{}_5S_x}{S(20)} = \frac{\int_x^{x+5} e^{-ra} l(a) \int_s^\omega e^{-ry} f(y) l(y) l(y+a) / l(y-0.75) dy da}{\int_x^{x+5} e^{-ra} l(a) da \int_s^\omega e^{-ry} f(y) l(y) l(y+20) / l(y-0.75) dy} \quad (4)$$

where $f(y)$ represents the age-specific fertility schedule, and $l(a)$ the life table survivorship, of men rather than of women and the ages between which childbearing occurs $s$ and $\omega$ are also those of men.

Equations 3 and 4 can be evaluated numerically using model life tables and fertility schedules and different age structures. Then a regression model that predicts life table survivorship can be fitted to these simulated data on parental survival. The estimation equation used for maternal orphanhood after age 20 is analogous to those proposed for orphanhood since birth (Timæus 1991, 1992). It is based on the observation that the proportion of respondents with living mothers equals a life table survivorship ratio, $l(B+N)/l(B)$, where $N$ is the age of the respondents and $B$ lies close to the mean age at childbearing, but also depends on $N$ (Brass and Hill 1973). For practical applications, however, it is more convenient to adjust the proportions slightly on the basis of the mean age at childbearing and to estimate survivorship for a rounded base age, $b$, close to $B$, and a duration of exposure, $n$, which is a multiple of five years. Moreover, for orphanhood after age 20 years, exposure starts 20 years after $B$. Thus, survivorship is estimated from a base age of 45 years and the equation used to make the estimates takes the form:

$${}_{n-20}p_{45} = a(n) + b(n)\bar{M}^f + c(n)\,{}_5S_{n-5}(\tau)/S(20,\bar{t}).$$

The equivalent equation sometimes gives poor results for men. More accurate estimates can be obtained if information on the survival of fathers in two adjoining age groups is used to infer mortality. If age patterns of mortality and childbearing differ from the average patterns reflected in the regression coefficients, the proportions of respondents with a living father in the upper age group and in the lower age group are shifted in compensating directions in comparison to the proportion at the age dividing the two groups (Timæus 1992). If one estimates life table measures from data on two age groups for a duration of exposure equal to their midpoint, one reduces the sensitivity of the results to variation in the slope of the relationship between parental survival and life table survivorship. The mean age of childbearing of men in developing countries averages a little less than 35 years. Thus, survivorship ratios can be estimated from orphanhood after age 20 by using a base age of 55 years and a model of the form:

$$
\begin{aligned}
{}_{n-20}p_{55} &= a(n) + b(n)\overline{M}^{m} \\
&\quad + c(n)\,{}_{5}S_{n-5}(\tau)/S(20,\overline{t}) \\
&\quad + d(n)\,{}_{5}S_{n}(\tau)/S(20,\overline{t}) \ .
\end{aligned}
$$

The coefficients for the different age groups defined by $n$ are presented in Table 26.1 and Table 26.2.

## EXTENSIONS OF THE METHOD

Chackiel and Orellana (1985) point out that, in addition to analysing orphanhood data from two inquiries using methods for synthetic cohorts, one can collect data in a single inquiry that can be used to produce up-to-date estimates in the same way. What is required in addition to the usual items about parental survival is information on the dates when parents died. For example, the inquiry might ask about the year and month when the parent died or how many years ago they died. If the dates when parents died are reported with reasonable accuracy, this information can be used to reconstruct the proportion of respondents who had living parents five and ten years earlier. From these successive cross-sections, one can construct synthetic cohort measures of parental survival that are formally identical to those generated from data collected in a series of separate inquiries. Thus, they can be analysed using the procedure for estimating mortality from orphanhood in adulthood that is described here with reference to data from multiple inquiries.

Rather few inquiries have tried to collect information on when parents died. In some of them the quality of the responses has been very poor, but in other inquiries the reported dates of occurrence of deaths in the previous decade or so, which are the deaths that are of most analytic interest, seem to have been quite well reported.

An alternative way of distinguishing between more recent and more distant death of parents that may yield better quality data is to ask whether parents died before or after some other important demographic event in the respondents' past, such as getting married or becoming a parent. Methods for making estimates of mortality from data of this type are described alongside other methods for the analysis of orphanhood data from a single inquiry.

## FURTHER READING AND REFERENCES

The basic orphanhood method is discussed in all the classic manuals on indirect estimation (Sloggett, Brass, Eldridge *et al.* 1994; UN Population Division 1983) but, with the exception of the United Nations manual on estimating adult mortality (United Nations Population Division 2002), these manuals give emphasis to the older variant of the method that uses weighting factors to produce life table indices, rather than the regression-based method normally used today. Although regression-based methods for women had been proposed previously (Hill and Trussell 1977; Palloni and Heligman 1985), regression methods for estimating men's mortality were first developed by Timæus (1992). His article also surveys earlier contributions to the literature and discusses the theoretical basis of the method.

Methods for constructing parental survival data for synthetic cohorts and estimating adult mortality from them were first proposed in the 1980s (Chackiel and Orellana 1985; Timæus 1986; UN Population Division 1983; Zlotnik and Hill 1981). The version of this approach that focuses on orphanhood after age 20 and is described here was first proposed by Timæus (1991).


Blacker JGC. 1984. "Experiences in the use of special mortality questions in multi-purpose surveys: the single-round approach," in *Data Bases for Mortality Measurement*. New York: United Nations, pp. 79–89. http://www.un.org/esa/population/publications/UN_1984_Data_Bases_for_Mortality_Measurement/UN1984_Data_Bases.htm

Blacker JGC and JM Gapere. 1988. "The indirect measurement of adult mortality in Africa: results and prospects," in *African Population Conference, Dakar, 1988*. Liège: International Union for the Scientific Study of Population, Vol. 2:3.2.23–38.

Brass W and K Hill. 1973. "Estimating adult mortality from orphanhood," in *International Population Conference, Liège, 1973*. Liège: International Union for the Scientific Study of Population, Vol. 3:111–123.

Chackiel J and H Orellana. 1985. "Adult female mortality trends from retrospective questions about maternal orphanhood included in censuses and surveys," in *International Population Conference, Florence, 1985*. Liège: International Union for the Scientific Study of Population, Vol. 4:39–51.

Hill K. 1984. "An evaluation of indirect methods for estimating mortality," in Vallin, J, Pollard John H and L Heligman (eds). *Methodologies for the Collection and Analysis of Mortality Data*. Liège: Ordina, pp. 145–176.

Hill K and TJ Trussell. 1977. "Further developments in indirect mortality estimation", *Population Studies* **31**(2):313–334. doi: 10.1080/00324728.1977.10410432

Palloni A and L Heligman. 1985. "Re-estimation of structural parameters to obtain estimates of mortality in developing countries", *Population Bulletin of The United Nations* **18**:10–33.

Palloni A, M Massagli and J Marcotte. 1984. "Estimating adult mortality with maternal orphanhood data: analysis of sensitivity of the techniques", *Population Studies* **38**(2):255–279. doi: 10.1080/00324728.1984.10410289

Preston SH and N Chen. 1984. *Two Census Orphanhood Methods for Estimating Adult Mortality, with Applications to Latin America*.

Preston SH and AJ Coale. 1982. "Age structure, growth, attrition and accession: A new synthesis", *Population Index* **48**(2):217–259.

Sloggett A, W Brass, SM Eldridge, IM Timæus, P Ward and B Zaba. 1994. *Estimation of Demographic Parameters from Census Data.* Tokyo, Japan: United Nations Statistical Institute for Asia and the Pacific.

Timæus I. 1986. "An assessment of methods for estimating adult mortality from two sets of data on maternal orphanhood", *Demography* **23**(3):435–450. doi: 10.2307/2061440

Timæus IM. 1991. "Estimation of mortality from orphanhood in adulthood", *Demography* **28**(2):213–227. doi: 10.2307/2061276

Timæus IM. 1992. "Estimation of adult mortality from paternal orphanhood: a reassessment and a new approach", *Population Bulletin of the United Nations* **33**:47–63.

UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

UN Population Division. 2002. *Methods for Estimating Adult Mortality*. New York: United Nations, Department of Economic and Social Affairs, ESA/P/WP.175. http://www.un.org/esa/population/techcoop/DemEst/methods_adultmort/methods_adultmort.html

Zlotnik H and K Hill. 1981. "The use of hypothetical cohorts in estimating demographic parameters under conditions of changing fertility and mortality", *Demography* **18**(1):103–122. doi: 10.2307/2061052

# Survey Data and Direct Methods

# Chapter 27 Estimation of adult mortality from sibling histories

*Ian M Timæus*

## DESCRIPTION OF METHOD

This method calculates adult mortality directly from data supplied by adults on their siblings (that is brothers and sisters). It can only be applied when an inquiry has collected full sibling histories. Such histories ask each respondent for the name, sex, age, survival status and, if dead, age at and year of death of each of their siblings born to the same mother. Information on brothers is used to estimate the mortality of men and information on sisters to estimate the mortality of women. Many surveys only collect sibling histories from women, but sibling histories collected from male respondents can be analysed using exactly the same methods. As respondents and their siblings are about the same age, on average, sibling histories can be used to measure mortality over approximately the same range of ages as the ages of respondents that the histories are collected from.

Collecting sibling histories is a complex process that requires careful training and supervision of field staff to be executed correctly. It is not an appropriate methodology to use in a census. Many Demographic and Health Surveys (DHS) collect sibling histories (referred to by the DHS as the 'Maternal Mortality Module'). While most of these surveys have only collected histories from women aged 15 to 49, as it is this group of women who complete a detailed individual interview, some DHS have collected sibling histories from men as well.

One advantage that sibling methods have over questions about household deaths is that only censuses or unusually large surveys can capture information on enough deaths in households in the year before the inquiry to yield mortality estimates that are sufficiently precise to be useful. Because respondents report on several siblings, on average, and one can calculate rates based on several years of exposure, estimates can be made from sibling history data in smaller inquiries. Nevertheless, all methods for the estimation of adult mortality require data on several thousand households. Another advantage of the method is that the estimation procedure makes few assumptions and, in particular, does not assume that the population being studied is closed to migration. However, the results from the method will not be representative for small states or sub-national areas in which a substantial proportion of the population are in-migrants or have emigrated.

## BACKGROUND

The initial methods developed for estimating mortality from information on the survival of siblings were indirect methods based on the idea that the average age of the siblings of respondents of any age is close to the age of the respondents. The proportion of a respondent's siblings who are still alive is, therefore, a good estimator of life table survivorship to the age of the respondent (Hill and Trussell 1977; UN Population Division 1983). Unfortunately, field experience of this approach demonstrated that the quality of the data collected on siblings was often low because siblings who died before or shortly after the respondent's own birth were often omitted by respondents, who may not know about them at all (Blacker and Brass 1983; Zaba 1986).

Interest in estimating mortality from data on siblings was revived by the development of the sisterhood method for measuring maternal mortality (Graham, Brass and Snow 1989). This requires data on how many sisters of the respondent survived to the age of 15, how many of them died thereafter, and whether sisters who died were pregnant at the time of death or had been pregnant during the 6 to 8 weeks before death. Limiting the consideration of siblings to only those who survived to age 15 years excludes siblings who died while still young and, therefore, may have been

unknown to or forgotten by the respondent. The responses supplied to the first two of these questions by respondents in each five-year age group allow one to calculate the proportions still alive of sisters who survived to age 15. The indirect adult sibling method was subsequently developed so that the all-cause mortality of adult women could be estimated from these proportions still alive. Comparable data on respondents' brothers can be used to estimate the mortality of men.

The method for estimating adult mortality by calculating life tables directly from full sibling histories collected from adult respondents was pioneered by the Demographic and Health Surveys programme based on this earlier research (Rutenberg and Sullivan 1991). It is more ambitious as to how much information it collects from the respondents and makes more demands on them and the field staff conducting the interviews. However, by replacing indirect estimation based on models of demographic relationships with the direct measurement of adult death rates, it reduces the number of assumptions involved in producing the estimates and, more importantly, allows one to separate deaths in the recent past from more distant ones, which may be reported on less accurately.

## DATA REQUIREMENTS AND ASSUMPTIONS

### Tabulations of data required

The calculation of mortality rates directly from sibling history data involves much the same steps and decisions as the more familiar process of calculating child mortality rates from birth history data – indeed, the history of a respondent's full set of siblings, including the respondent, *is* the mother's birth history. However, in comparison with mortality estimates made for children, estimates for adults have very large sampling errors. This reflects the facts that death rates are very much lower in adulthood than childhood and that, in a growing population, the number of siblings exposed to risk is small relative to the number of children reported on by mothers. In any household survey, therefore, far fewer sibling deaths than child deaths will be reported.

The calculation of cohort measures from sibling history data has little to recommend it either analytically or computationally, particularly given the ease with which modern survey analysis software can deal with the calculation of exposure times. Thus, this document focuses on the calculation of age-period death rates from sibling history data and on deriving other mortality indicators from them.

To calculate women's mortality one needs to tabulate:

- The number of deaths of respondents' sisters by time period and five-year age group of the sisters at the time of their death.
- The number of sister-years of exposure by time period and five-year age group of the sister at the time of exposure.

To calculate men's mortality one needs to tabulate:

- The number of deaths of respondents' brothers by time period and five-year age group of the brothers at the time of their death.
- The number of brother-years of exposure by time period and five-year age group of the brother at the time of exposure.

The calculations of exposure time should usually *exclude* the respondent himself or herself (and do so by definition if one is analysing sibling histories for the opposite sex). This requirement is explained in the discussion of the important assumptions made by the method.

Tables on respondents' own-sex siblings (i.e. women's sisters and men's brothers) should be weighted only by any sample or design weights provided with the data. Tables on respondents' opposite sex siblings (i.e. women's brothers and men's sisters) should be further weighted by the inverse of the number of surviving own-sex siblings of the individual respondent making the reports. This requirement is also explained in the discussion of the important assumptions made by the method.

The time period over which exposure is measured can be defined either in terms of calendar dates or relative to the date when the respondent was interviewed. The latter approach makes efficient use of the data as it ensures that the experience of respondents' siblings during the incomplete years in which the interviews occur are included in the analysis. Calculating death rates for particular calendar years, however, has the advantage of yielding precisely time-referenced results that can be compared with those from other sources and is to be preferred. This approach usually entails discarding the data on the year in which the histories are collected but, if the majority of interviews took place late in that year, one might opt to include it in the most recent period.

The workbook (see website) is set up to calculate death rates for two successive periods prior to the collection of the data. Many DHS survey reports present rates for the seven-year period preceding their collection (that is, 0–6 completed years) and one way to use the spreadsheet

would be to calculate rates for the three years prior to the inquiry and the four years prior to that. Alternately, death rates could be calculated for two four-year periods and for an eight-year period preceding the collection of the sibling histories. Experience suggests that the completeness of the reporting of dead siblings and accuracy with which their ages and dates of death are recalled often deteriorate rapidly for events occurring longer ago than that. Moreover, working with four-year periods rather than five-year ones minimizes errors that result from rounding of dates of death to five and ten years before the collection of the data.

The age groups of siblings for which the data are tabulated should broadly correspond to the ages of the respondents that the data were collected from. For example, in order to measure mortality between ages 15 and 60, one should ideally collect sibling histories from respondents aged 15 to 59. If data are only collected from adults, few of them have siblings that are still young children and, even if reports on children are not biased by failure to report some dead siblings, the estimates are likely to have large sampling errors and will not be representative of all young children. Thus, for DHS and other surveys that collect sibling histories from respondents aged 15 to 49, the preferred summary index of mortality is ${}_{35}q_{15}$, the probability that someone aged 15 dies before their 50th birthday.

Although the conditional probability of dying between ages 15 and 60 (${}_{45}q_{15}$) is widely used by international agencies and other organizations as their preferred summary index of adult mortality, the number of deaths in the 55–59 year age group reported by respondents aged 15 to 49 will be small relative to those reported for younger age groups. Thus, rather than calculating ${}_{45}q_{15}$ directly, it is better to do so by fitting a model life table to the estimates for ages 15 to 54 and extrapolating in this model to obtain mortality in the final five-year age group. This approach is implemented in the workbook (see website).

To eliminate ambiguities related to polygynous marriage and to remarriage, interviewers in most inquiries are instructed that 'siblings' means children born to the same mother. Whether or not this has been done, the reports should usually be accepted as they are. So long as respondents have the same group of relatives in mind when they are listing dead siblings as when they are listing those who are still alive, it is immaterial for the purpose of estimating mortality exactly who the parents of the siblings are.

If sibling histories have been collected from both men and women, their responses should usually be tabulated separately so that the two sets of data can be weighted appropriately and checked against each other.

**Important assumptions**

An inherent limitation of sibling-based methods for measuring adult mortality is that they underestimate mortality insofar as mortality clusters within sibships (i.e. sets of brothers and/or sisters born to the same mother). Clustering occurs whenever deaths are more concentrated in a small proportion of sibships than would be expected by chance and results from between sibship heterogeneity in individuals' risk of dying (Zaba and David 1996). It causes downward bias in the mortality estimates simply because fewer members of a high mortality sibship than a low mortality sibship of the same size remain alive to answer questions about their siblings. It is impossible to correct fully for this because, at the extreme, sets of siblings whose members have all died are not reported on at all. Although one can assume something about them, there is no way of determining empirically from data collected retrospectively how many of these sibships existed or what their sizes were.

Estimates of the age pattern and trend in mortality will be biased if the extent to which mortality clusters within sibships varies with age. For example, if characteristics shared by sibs (e.g. genetic factors, early-life experiences, socio-economic status, life styles, and location) strongly influence the mortality of middle-aged adults, whereas mortality before age 40 has a large random component, estimates for older adults will underestimate mortality by more than those for younger adults, producing a spurious impression of mortality increase over time.

The issue of bias related to multiple reporting of siblings has received substantial attention in the literature. The problem exists in survey as well as census data because the more times an individual would be reported in a census, the more likely they are to be have a sibling who reports on them included in a probability sample.[1] Moreover, even in surveys, potential exists for multiple responses about the

1 Note that the issue of calculating the sampling error of the mortality estimates correctly in surveys in which some individuals are reported on more than once and mortality is clustered at the level of the sibship, is distinct from the issue being discussed here, which is that of *bias* in the central estimate.

same individual. For example, if two daughters of the same mother are interviewed in the same household, there will be multiple reports about other members of the sibship. The standard approach to analysis used, for example, in DHS reports is based on the events and exposure time of reported siblings, leaving out the exposure time of the (surviving) respondent herself. Events and exposure time are weighted only by the respondent's sample weight, not taking into account numbers of surviving potential respondents in the sibship.

Trussell and Rodriguez (1990) demonstrate mathematically that for groups of sibships with an identical underlying risk of dying, this standard approach yields unbiased estimates of mortality. In effect, the reduction in the number of deaths reported in the numerator that occurs because dead people cannot report on one another and the exclusion of the exposure time of the living respondents from the denominator offset each other precisely to give the correct mortality rates for the sibships as a group.

The issue of the biases that could result from differential mortality by sibship size is bound up with the issue of multiple-reporting bias. It has attracted a lot of research interest because, unlike other factors that affect risk within sibships classified by sex and age of the respondent, each respondent's sibship size is known. If mortality does not vary with sibship size, the standard estimates are the same for every size of sibship, including one-person sibships that are excluded from the analysis because the respondent has nobody to report on, as well as for the population as a whole. Even if mortality varies by sibship size, the standard estimates remain unbiased for each sibship size, as pointed out by Masquelier (2013). To obtain mortality estimates for the population though, one must reweight the estimates for sibships of different sizes by the prevalence of sibships of that size in the population. When respondents are reporting on their own sex, one can achieve this by dividing the proportion of respondents from surviving sibships of each size by the estimated probability of surviving from the age at which siblings are counted as entering exposure to risk to the current age group of the respondents across all sibships of the same size. To do this for single-person sibships, their mortality has to be estimated by extrapolation from mortality in larger sibships.

Gakidou and King (2006) argue that, instead, sibships should include the exposure of the surviving respondent but should always be weighted in addition by the likelihood that they will be reported – that is, by the inverse of the number of potential respondents in the sibship. As in Masquelier's approach, an adjustment also must be made for sibships that go unreported because no member remains alive. In a multi-survey analysis of DHS full sibling histories, Obermeyer, Rajaratnam, Park *et al*. (2010) estimate that the effect of not adjusting for the likelihood of reporting can bias overall mortality estimates downward by as much as 20 per cent.

Masquelier (2013), however, argues that Obermeyer and her co-authors reweighted their data files inappropriately and, as a result, exaggerated the size of any bias. He emphasizes that, if one is going to reweight, it is important only to adjust for multiple reporting by siblings who survived to the initial age from which mortality is being measured. In addition though, he questions whether the observed variation in mortality by sibship size is necessarily real. Instead, he argues, it may be an artefact of greater omission of dead siblings in the histories reported for large sibships. Masquelier therefore recommends using the standard approach, without attempting to reweight the data on each sibship. Then, either mortality should be estimated for each size of sibship and a reweighted estimate for the population obtained in the way described a few paragraphs previously or the estimates should not be reweighted at all. As making separate estimates for each size of sibship requires either a very large sample survey or fitting models to the data by regression methods in order to smooth them, the latter approach is adopted here.

When histories are being analysed on siblings of the opposite sex (for example, histories collected from women concerning their brothers), the issues are rather different. In this case, the respondent is not a member of the group that is exposed to the risk of dying. However, the standard calculation will still give biased results for the population as a whole if the mortality of siblings of one sex is associated with the number of siblings of the opposite sex that report on them. Thus, for reports on the opposite sex a clear case exists for weighting each report by the inverse of the respondent's number of surviving siblings of their own sex as suggested by Gakidou and King (2006). Of course, questions about siblings of the opposite sex cannot generate any information on those sibships whose members have no living siblings of the respondent's sex. Thus, adopting this approach is equivalent to assuming that the mortality of individuals in such sibships is the same as the mortality of the rest of the population. In surveys that collect data

from both sexes, each sex supplies this information for the other and one can further weight the deaths and exposure reported by respondents by the inverse of the probability that siblings in each age group get reported on at all.

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

The initial step in the analysis of sibling history data should be to assess the extent of non-reporting and incomplete reporting in the data set, in particular how many respondents stated that they did not know whether a particular sibling remains alive or simply failed to answer the question. It is also important to assess the extent to which the siblings' dates of birth and ages, and their ages at and dates of death, are missing or have been imputed. If a lot of respondents failed to respond to these questions, the data supplied by those respondents who did answer them may not be representative of the population as a whole. Moreover, a high level of non-response may indicate that either the field staff or the respondents were having difficulty with the questions and it can be illuminating to determine whether the problem is concentrated among a minority of field staff or certain type of respondent. Heaping of reported dates and ages on particular ages and years also indicates that reporting is not very accurate. If the quality of the age and date data seems particularly poor, one might obtain better results from analysing the data set using the indirect adult sibling method.

If both women and men have been asked the relevant questions, one useful check on the completeness of the data is to assess how many siblings of each sex are reported, on average, by respondents of the other sex and whether the reported sex ratio at birth changes markedly as either the age of the respondent or the time since the birth of the siblings increases. It is fairly common to find that one sex (usually men) reports fewer siblings, and in particular fewer dead siblings, than the other. In other surveys, men and women may report similar numbers of siblings of each sex (after adjusting for the fact that respondents do not report on themselves) but that different numbers of them remain alive. The first type of discrepancy might result from differential age misreporting, but the second cannot.

Any bias due to clustering of mortality within families results in underestimates. Moreover, it seems unlikely that respondents invent siblings or report that their living siblings have died. Thus, the analysis should probably focus on the data supplied by respondents of the sex that reports most siblings and most siblings that have died.

## CAVEATS AND WARNINGS

- The only sibling methods that can be recommended produce probabilities of surviving from age 15 to ages later in adulthood conditional on being alive at exact age 15. In theory, it is possible to collect and analyse data on the deaths of siblings as children. Unfortunately, such reports are often very incomplete, particularly for siblings who died before or soon after the birth of the respondent. Thus, most applications of the method only attempt to measure siblings' mortality at age 15 or more. To produce a complete life table, one has to estimate survivorship from birth to age 15 using another source of data.
- Even in a large survey, the number of siblings that die each year in each age group is small. In most applications of the method, deaths and exposure need to be aggregated over several years in order to estimate death rates that are precise enough to be useful. Thus, the method is unlikely to be useful for detecting abrupt changes or fluctuations in adult mortality.
- As one works backward from the time of the survey, the number of siblings exposed to risk diminishes rapidly, particularly at older ages. Moreover, respondents tend to omit some dead siblings from their reports, particularly deaths that occurred quite a long time ago. Thus, the histories tend to underestimate mortality and this bias often gets worse for estimates that are more distant from the time of the survey. Thus, the direct sibling method should not be used to estimate death rates for more than 10 years before the data were collected. Often only the data on the last seven years are analysed.
- Given that the likelihood that dead siblings are omitted from the reports rises as the time since their death increases, the purpose of calculating rates for two periods before the survey is largely diagnostic. It enables the analyst to check whether the data indicate an implausible rise in mortality. One should be cautious about inferring trends in adult mortality from the internal evidence provided by a single set of sibling histories.
- The direct procedure for estimating adult mortality from information on siblings does not involve the assumption that the population is closed to migration. Nevertheless, it can be difficult to interpret sibling-based estimates of adult mortality for sub-national geographic units, such as

urban and rural areas or districts, or for respondents with particular socio-economic characteristics. This is because, although siblings usually share the same ethnic identity, many of the respondents' siblings will live in different places from the respondents themselves and their socio-economic characteristics may differ from those of the respondents. Estimates for sub-national populations are also likely to have very large sampling errors.

## APPLICATION OF METHOD

The procedure for estimating death rates from sibling history data is identical no matter whether one is analysing data on brothers, sisters, or siblings of both sexes and irrespective of whether the respondents are men, women, or both the sexes. The workbook (see website) is set up to calculate death rates for both brothers and sisters for two periods of time preceding an inquiry and for the entire period covered by them combined. Separate worksheets are provided for the analysis of data provided by male and female respondents. The workbook can produce estimates for periods of any length and for data tabulated by 'years before survey' or for calendar-year periods corresponding broadly to them.

Two tables are required in order to calculate the death rates for siblings of each sex reported on by respondents of a particular sex, which amounts to potentially four pairs of tables. One table should contain counts of deaths of siblings by year and age and the other table should contain person-years of observation of siblings exposed to the risk of death by year and age. The data can be tabulated for the age groups and periods that are going to be used in the analysis. Alternatively, one could produce the tables for single years of age and time so that the counts can be can be aggregated over either dimension into any set of wider intervals that is subsequently found to be of interest.

Assuming that data set lacks exact dates of birth for some siblings and exact dates of death for some siblings that have died, the most satisfactory way of addressing this limitation of the data is to impute exact dates using random numbers to place respondents within the range of dates at which the event could have occurred (Stanton, Abderrahim and Hill 1997). Survey organizations such as MeasureDHS may have done this before distributing the data from surveys that they conducted.

For imputation and analysis of the data on particular siblings to be possible, one needs to know either their year of birth or their current age, if they are alive, together with a year of death, age at death or time in years since death, if they are dead. If both the date of birth and date of death are incomplete, one would generally randomly assign the person an exact date of birth before assigning them a consistent date of death. Care should be taken to record the seed for the random number generator used for this, so that the imputed dates can be reproduced exactly if the need arises to recreate the data files being used to estimate mortality from the original data.

The details of the procedure that should be used to impute exact dates depends on precisely what questions were asked in the sibling histories about ages and times of death. A few examples will suffice to illustrate the principles involved in the calculations. If the respondent was interviewed on 23/11/2011 and reported that one of their siblings was 33 years old, that person's date of birth must fall on or between 24/11/1977 and 23/11/1978. If a sibling is reported to have been born in October 1972 and to have died at age 17, one would first randomly assign them an exact date of birth, perhaps 14/10/1972, and then assign them an exact date of death on or between 14/10/1989 and 13/10/1990. If the respondent also reported a year of death for the sibling, the range of dates within which one randomly chooses a date of death should be restricted to the correct year.

A little care is needed to ensure that no siblings whose age at death equals the number of years since their birth are assigned a date of death later than the respondent's date of interview. For example, if the respondent was interviewed on 28/2/2003 and reported that their sibling was born in 1980 and died at age 23, then that person must have been born in the first two months of 1980 and have died on or after that date in 2003, but before the end of February. The imputation procedure should also ensure that the imputed dates of birth occur in the correct temporal order of birth.

Once every sibling has been assigned an exact date of birth and, if they have died, of death, it is straightforward to identify the age groups in which deaths occurred and to divide the person's life up between the age groups and periods (Stanton, Abderrahim and Hill 1997). Modern survey analysis software often contains commands that semi-automate this process.

Graphically, mortality is measured for the age group and period of time defined by the heavy lines in Figure 27.1. An individual's life course by age and period is represented by the diagonal lines (as with a conventional Lexis diagram). The age group for which mortality is to be calculated is

defined to have a lower bound of $x_l$, and an upper bound of $x_h$. The period of time for which mortality is to be calculated is defined as $(t_2 - t_1)$, where $t_2$ is its end date and $t_1$ its start date. Thus, any person aged $x$ at $t_1$ who does not die before $t_2$ will be aged $x_{t2} = x_{t1} + (t_2 - t_1)$ at time $t_2$. For statistics on adult mortality, both age and calendar time are almost always measured in years.

Note that each sibling's exposure in any year is almost always divided between two adjacent ages. Five possible scenarios are portrayed on Figure 27.1, labelled (a) through (e). Denoting the age at death of individuals dying in this age group and period as $x_d$, the contribution that each scenario makes to the person-years of exposure of the respondents' siblings in this age group and period can be determined by the calculations shown in Table 27.1. Using these rules, one can calculate the exposure in the age group and period of each sibling reported on in the sibling histories. Summing exposure across all siblings gives total years of exposure to risk in the age group during the period, the denominator for the death rate. Summing the deaths occurring in the same age range and period provides the numerator for the rate.

Once the tables of deaths and exposure have been produced, various measures of mortality can be produced using standard life table calculations. These calculations are carried out for data on five-year age groups in the workbook (see website). The age-specific death rate, ${}_5M_x$, is calculated by dividing the deaths in a five-year age group in a specific year or period of years by the person-years spent exposed to the risk of dying in that age group during that period:

$$ {}_5M_x = \frac{{}_5D_x}{{}_5N_x}. $$

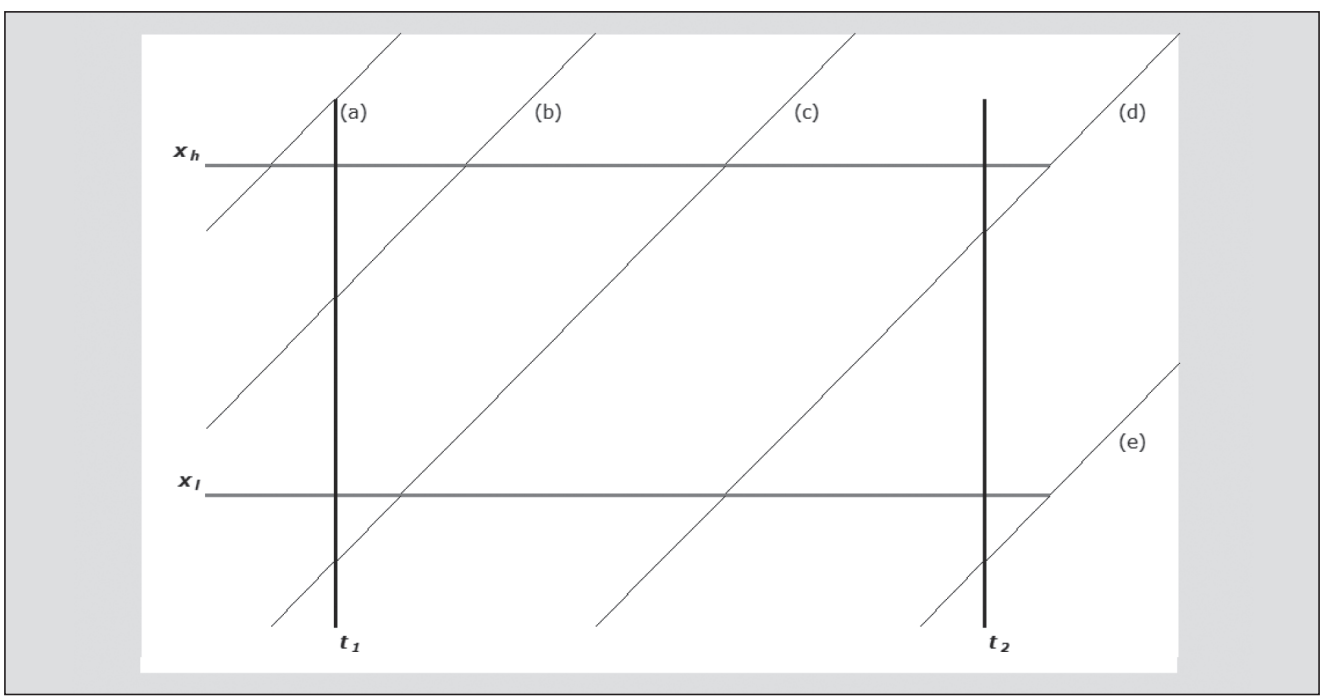
The probability of dying in a five-year age group, ${}_5q_x$, in the years concerned can be calculated from the corresponding death rate using the standard formula, which assumes that the deaths are evenly distributed across the age group:

$$ {}_5q_x = \frac{5 \times {}_5M_x}{\left(1 + 2.5 \times {}_5M_x\right)}. $$

The probability of surviving a five-year age group, ${}_5p_x$, is $1 - {}_5q_x$.

From the series of estimates of ${}_5p_x$, one can calculate the cumulative probability of dying between age 15 and age 50 for the period, ${}_{35}q_{15}$, by multiplying together the intermediate five-year probabilities of surviving to obtain the

***FIGURE 27.1*** *Lexis diagram showing calculation of exposure to risk*



probability of surviving from 15 to 50, ${}_{35}q_{15}$, and subtracting this probability from 1 to obtain its complement:

$$ {}_{35}q_{15} = 1 - {}_5p_{15} \times {}_5p_{20} \times {}_5p_{25} \ldots \times {}_5p_{45} \,. $$

The 95 per cent confidence intervals of these summary measures of adult mortality provided in the workbook (see website) are calculated using Greenwood's formula. This formula assumes that the data are generated from a simple random sample and so will overstate the precision of indices based on data from cluster surveys.

The workbook (see website) produces plots of the logits of the conditional probabilities of surviving from age 15 to each higher age against the equivalent values of logit survivorship in a standard life table. Such plots are useful for evaluation of the quality of the estimates. Data errors usually show up as irregularities in the series or in the form of downward curvature of the series in the oldest age groups. The latter pattern is indicative of underestimation of mortality due to exaggeration of ages and ages at death.

Finally, the workbook (see website) fits a 2-parameter relational model life table to the series of ${}_np_{15}$ values by means of a simple linear regression across the entire age range 15 to 55 years. This smoothes out some of the errors in the series. Fitted values of ${}_{35}q_{15}$ and ${}_{45}q_{15}$ are extracted from this life table. The workbook can fit the model life table and calculate these mortality indices using either a standard from the General family of United Nations model life tables (UN Population Division 1982) or one from any of the four families of Princeton model life tables (Coale, Demeny and Vaughan 1983). The standard life table should be chosen to have an age pattern of mortality within adulthood that resembles that of the population being studied. Another life table can be used as a standard if there is reason to believe that it resembles more closely the pattern of adult mortality in the population being studied. The most suitable life table may not be from the family of models that best captures the relationship between child and adult mortality. If nothing is known about the age pattern of mortality in adulthood, use of the United Nations General or Princeton West models is recommended.

## WORKED EXAMPLE

These calculations are illustrated in Table 27.2 using data collected from female respondents about their sisters in the 2001 Maternal Mortality Survey of Bangladesh (available on the MeasureDHS website). Note that this is an unusually large survey. The deaths and exposure of the respondents' sisters in each five-year age group have been cumulated across the seven-year period preceding the survey. After the calculation of the death rates and estimates of life table survivorship, the latter have been smoothed by fitting a 2-parameter relational model life table using a Princeton South model life table as the standard.

Plots of the estimates against the standard are shown for the sisters, distinguishing between the periods 0 to 2 years and 3 to 6 years before the survey, in Figure 27.2. The equivalent plots for the respondents' brothers are shown in Figure 27.3. For these particular data, the smoothed estimates of ${}_{35}q_{15}$ and ${}_{45}q_{15}$ are almost identical to the estimates calculated directly from the data at 90 per 1000 and 167 per thousand respectively.

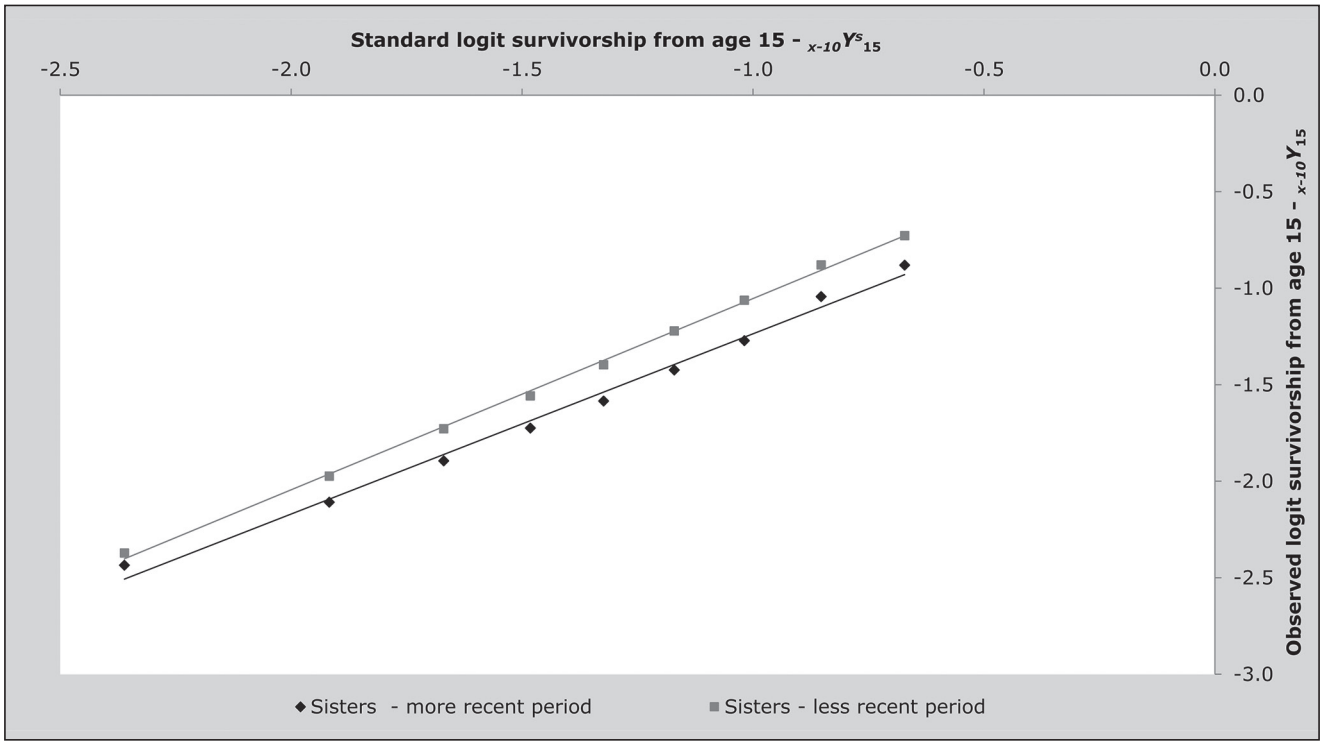
*TABLE 27.1 Algorithm for determining exposure to risk*

| Scenario | Description | Defining rule(s) | Exposure of survivors during the period | Exposure of decedents (if death occurs during the period) |
|---|---|---|---|---|
| (a) | Aged older than $x_h$ at $t_1$ | $x_{t1} > x_h$ | 0 | 0 |
| (b) | Aged between $x_l$ and $x_h$ at $t_1$. Attains $x_h$ in the period | $x_l < x_{t1} < x_h$<br>$x_{t1} + (t_2 - t_1) > x_h$ | $x_h - x_{t1}$ | $x_d - x_{t1}$ |
| (c) | Attains $x_l$ and $x_h$ in the period | $x_l > x_{t1}$<br>$x_{t1} + (t_2 - t_1) > x_h$ | $x_h - x_l$ | $x_d - x_l$ |
| (d) | Attains $x_l$ in the period. Period ends before $x_h$ | $x_l > x_{t1}$<br>$x_l < x_{t1} + (t_2 - t_1) < x_h$ | $x_{t1} + (t_2 - t_1) > x_l$ | $x_d - x_l$ |
| (e) | Does not attain $x_l$ in the period | $x_{t1} + (t_2 - t_1) < x_l$ | 0 | 0 |

*TABLE 27.2 Direct calculation of age-specific death rates and the probabilities of dying between age 15 and ages 50 and 60, women, Bangladesh, 1994–2001*

| Age group $x$ to $x+4$ | Deaths of sisters | Person-years of exposure | Age-specific death rate ${}_5M_x$ | Five-year survivorship ${}_5p_x$ | Cumulative survivorship ${}_{x-10}p_{15}$ | Logits ${}_{x-10}Y_{15}$ | Smoothed logits ${}_{x-10}Y_{15}$ |
|---|---|---|---|---|---|---|---|
| 15–19 | 350.3 | 211,840.6 | 0.00165 | 0.9918 | 0.9918 | –2.3956 | –2.4318 |
| 20–24 | 436.8 | 241,208.5 | 0.00181 | 0.9910 | 0.9828 | –2.0235 | –2.0109 |
| 25–29 | 488.0 | 241,111.4 | 0.00202 | 0.9899 | 0.9729 | –1.7909 | –1.7758 |
| 30–34 | 455.0 | 210,963.3 | 0.00216 | 0.9893 | 0.9625 | –1.6225 | –1.5978 |
| 35–39 | 417.5 | 160,378.1 | 0.00260 | 0.9871 | 0.9500 | –1.4727 | –1.4472 |
| 40–44 | 377.5 | 97,268.6 | 0.00388 | 0.9808 | 0.9318 | –1.3072 | –1.3020 |
| 45–49 | 242.0 | 50,456.1 | 0.00480 | 0.9763 | 0.9097 | –1.1550 | –1.1581 |
| 50–54 | 169.4 | 19,621.2 | 0.00863 | 0.9577 | 0.8713 | –0.9561 | –1.0002 |
| 55–59 | 56.7 | 6,276.6 | 0.00904 | 0.9558 | 0.8328 | –0.8027 | –0.8286 |
| ${}_{35}q_{15}$ (95% CI) | | | 0.090 | (0.086–0.094) | | ($\alpha$ =–0.191) | 0.090 |
| ${}_{45}q_{15}$ (95% CI) | | | 0.167 | (0.155–0.179) | | ($\beta$ = 0.949) | 0.160 |

*FIGURE 27.2 Logit survivorship from age 15 plotted against a Princeton South model life table, women, Bangladesh, 1994–2001*



## DIAGNOSTICS, ANALYSIS AND INTERPRETATION

### Checks and validation

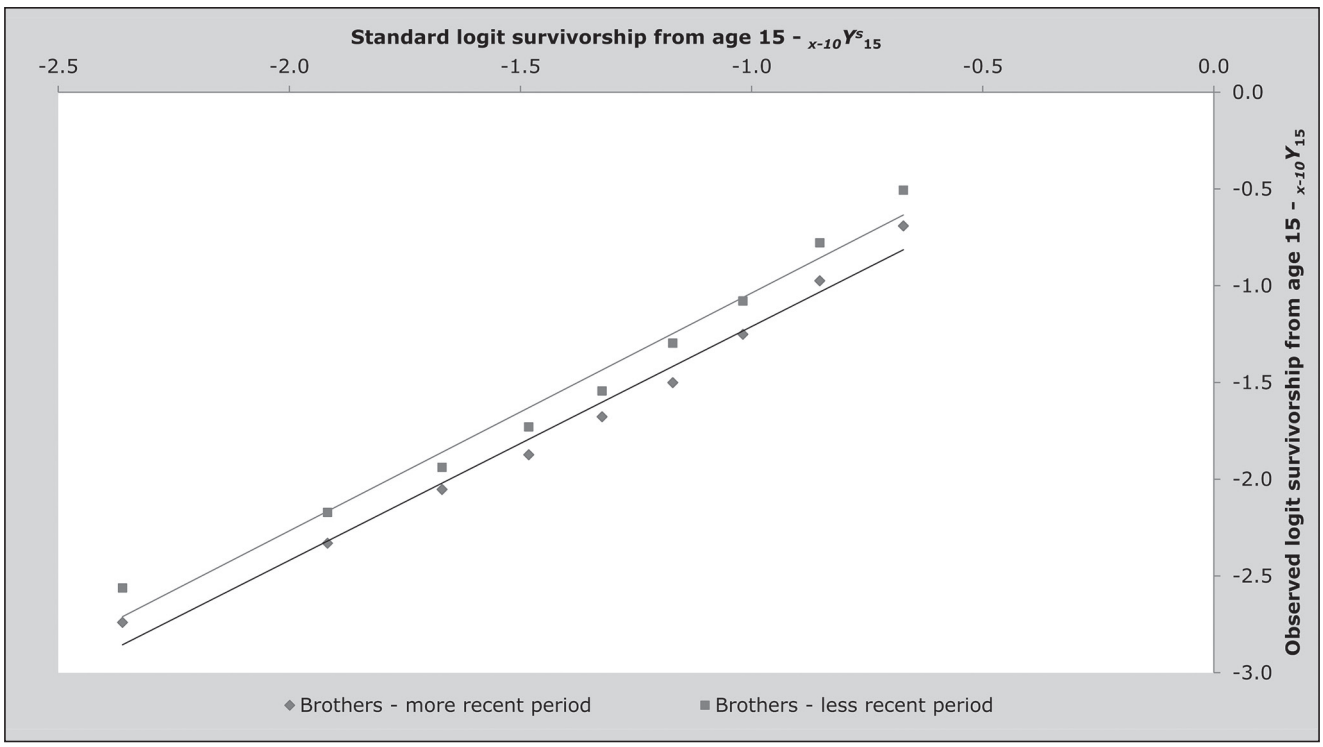
If sibling histories have been collected from both men and women in a census or a large survey, analysing them separately for male and female *respondents* can be recommended in order to compare the consistency of their reports. The mortality of individuals of a particular sex, as reported by their brothers, should be the same as the mortality of the same individuals as reported by their sisters. If it is not, this may indicate significant bias in the estimates for one or both sexes. While consistency of reporting does not guarantee accuracy, statistically significant differences between the estimates obtained from male or female respondents do imply that at least one sex, and possibly both of them, are answering the questions inaccurately. This check on the results cannot be carried out on the data from the 2001 Bangladesh Maternal Mortality Survey as sibling histories were collected only from women in this study.

### Interpretation

The results of the illustrative application of direct calculation of death rates from sibling histories using data from the 2001 Maternal Mortality Survey of Bangladesh are shown in Figures 27.2 and 27.3. They are encouraging. The plotted points are not at all erratic and do not curve away at older ages. There is some curvature in the series for men, particularly for the earlier period, but this spans the entire series of points from age 20 up to age 60, rather than affecting just the older ages. It probably indicates that the age pattern of mortality in this population differs from that in the standard life table.

The estimates for the two periods are consistent with each other for both sexes, in each case suggesting that adult mortality declined substantially during the 1990s. Thus, the probability of dying between ages 15 and 50 of men is estimated to have dropped from the 104 per thousand to 76 per thousand between the period 3 to 6 years before the survey and 0 to 2 years before the survey. The probability of dying between ages 15 and 50 of women dropped from 107 per thousand to 73 per thousand between the same two periods.

*FIGURE 27.3 Logit survivorship from age 15 plotted against a Princeton South model life table, men, Bangladesh, 1994–2001*



Although the overall probability of dying between ages 15 and 50 in Bangladesh is very similar for men and women and has declined at a similar rate for both sexes, Figures 27.2 and 27.3 reveal that Bangladeshi men and women have very different age patterns of mortality within adulthood. Mortality rises much more steeply with age for the men than the women. The $\beta$ parameter of the model life table fitted to the 1994 to 2001 data for women is 0.95 while, in the equivalent model life table for men, it is 1.14. Thus, if one examines the death rates for the five-year age groups, they show that women have higher mortality than men in Bangladesh at ages 15 to 40, but that men in their 40s and 50s have higher mortality than women.

The internal regularity of each of the four series of estimates from this survey in Bangladesh, the consistency of the estimates for the two periods before the survey, and the plausibility of the age pattern of mortality as assessed against external standards, all represent evidence that the method worked well in this survey. The most surprising feature of the results is the very large drop in adult mortality that they suggest occurred in Bangladesh in the second half of the 1990s.

### Performance in populations with generalized HIV epidemics

The HIV epidemic poses two problems for methods of estimating mortality based on the survival of relatives (UN Population Division 1982). First, both the sexual and vertical routes of transmission produce significant selection biases in data collected in surveys on the survival of relatives. Second, the incidence of HIV infection is concentrated among young adults. Thus, populations with significant AIDS mortality have very different age patterns of mortality from both other populations and existing systems of model life tables.

A major advantage of sibling methods of measuring adult mortality over questions about other relatives is that they are free of selection biases arising from direct transmission of the virus. Some residual bias due to clustering of AIDS mortality within sibships will remain. All the children born to a woman after she becomes infected are at risk of infection by vertical transmission. Moreover, the risk of HIV infection tends to vary markedly between localities and siblings often live close to each other. The impact of this, however, will be relatively small compared with the biases that affect data that parents have supplied about their children or vice versa. Moreover, direct estimates of mortality from sibling histories have an advantage over the indirect adult sibling method in populations with substantial AIDS mortality in that they measure the age pattern of mortality directly – nothing has to be assumed about it.

## EXTENSIONS AND VARIANTS OF THE METHOD

In order to extract the maximum useful information from sibling histories in the presence of both reporting and sampling errors, analysts have resorted to multi-country analyses of sibling-level data files using regression models to impose some discipline on the results (for example, Obermeyer, Rajaratnam, Park *et al.* 2010; Timæus and Jasseh 2004). For instance, Timæus and Jasseh incorporate a 2-parameter standard mortality schedule in their regression model of the log mortality rates in order to smooth the data. They allow the regression coefficient for the standard (which determines the age pattern of mortality) to vary between countries but not to change over time. Moreover, they assume that, while the speed of decline in mortality from causes other than AIDS varies between countries, it follows a log-linear trend in them all. Other analysts have constrained their estimates in different ways.

## FURTHER READING AND REFERENCES

The direct method of calculating adult mortality directly from sibling history data is not discussed in the classic manuals on indirect estimation. Although their report is focused primarily on measuring maternal mortality, Stanton, Abderrahim and Hill (1997) discuss a number of important issues relating to the estimation of all-cause mortality from full sibling histories in some detail, including the imputation of exact dates of birth and death and the calculation of exposure time. Biases related to differential mortality by family size and multiple reporting of siblings are discussed by Gakidou and King (2006), Masquelier (2013), and others.


Blacker JGC and W Brass. 1983. "Experience of retrospective enquiries to determine vital rates," in Moss, L and H Goldstein (eds). *The Recall Method in Social Surveys. London: University of London Institute of Education*, pp. 48–61.

Coale AJ, P Demeny and B Vaughan. 1983. *Regional Model Life Tables and Stable Populations*. London: Academic Press.

Gakidou E and G King. 2006. "Death by survey: estimating adult mortality without selection bias from sibling survival data", *Demography* **43**(3):569–585. doi: 10.1353/dem.2006.0024

Graham W, W Brass and RW Snow. 1989. "Estimating maternal mortality: The sisterhood method", *Studies in Family Planning* **20**(3):125–135. doi: 10.2307/1966567

Hill K and TJ Trussell. 1977. "Further developments in indirect mortality estimation", *Population Studies* **31**(2):313–334. doi: 10.2307/2173920

Masquelier B. 2013. "Adult mortality from sibling survival data: A reappraisal of selection biases?", *Demography* **50**(1):207–228. doi: 10.1007/s13524-012-0149-1

Obermeyer Z, JK Rajaratnam, CH Park, E Gakidou *et al*. 2010. "Measuring adult mortality using sibling survival: a new analytical method and new results for 44 countries, 1974-2006", *PLoS Medicine* **7**(4):e1000260. doi: 10.1371/journal.pmed.1000260

Rutenberg N and JM Sullivan. 1991. "Direct and indirect estimates of maternal mortality from the sisterhood method," Paper presented at Demographic and Health Surveys World Conference, August 5–7, 1991, Washington, DC Columbia. Macro International. Vol. 3:1669–1696.

Stanton C, N Abderrahim and K Hill. 1997. *DHS Maternal Mortality Indicators: An Assessment of Data Quality and Implications for Data Use.* Calverton: Macro International.

Timæus IM and M Jasseh. 2004. "Adult mortality in Sub-Saharan Africa: evidence from Demographic and Health Surveys", *Demography* **41**(4):757–772. doi: 10.1353/dem.2004.0037

Trussell J and G Rodriguez. 1990. "A note on the sisterhood estimator of maternal mortality", *Studies in Family Planning* **21**(6):344–346. doi: 10.2307/1966923

UN Population Division. 1982. *Model Life Tables for Developing Countries*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/77. http://www.un.org/esa/population/techcoop/DemMod/model_lifetabs/model_lifetabs.html

UN Population Division. 1983. *Manual X: Indirect Techniques for Demographic Estimation*. New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/81. http://www.un.org/esa/population/techcoop/DemEst/manual10/manual10.html

Zaba B. 1986. *Measurement of Emigration using Indirect Techniques: Manual for the Collection and Analysis of Data on Residence of Relatives.* Liège: Ordina.

Zaba B and PH David. 1996. "Fertility and the distribution of child mortality risk among women", *Population Studies* **50**(2):263–278. doi: 10.1080/0032472031000149346

# Maternal Mortality

# Chapter 28 Introduction to maternal mortality analysis

*Kenneth Hill*

Maternal mortality has long been a focus of national health services, but its salience has increased over the last two decades with the establishment of quantitative goals. During that time, the international community has periodically established targets for the reduction of maternal mortality, measured as the Maternal Mortality Ratio (MMR), maternal deaths per 100,000 live births. The World Summit for Children in 1990 set the goal of reducing MMR by half between 1990 and 2000. The 1994 International Conference on Population and Development (ICPD) reiterated this goal, but set the additional longer-term target of reducing the rate by a further half by 2015. The Millennium Summit in 2000 adopted the ICPD target for the fifth MDG (the improvement of maternal health). The target was thus to reduce the MMR by three-quarters between 1990 and 2015. The 2011 report of the Commission on Information and Accountability for Women's and Children's Health, established by the Secretary-General of the United Nations, reaffirmed the importance of timely reporting on MMR as one of 11 indicators of maternal, newborn, and child health. It is thus clear that the measurement of maternal mortality has a very high priority. This chapter discusses broad options for such measurement.

## DEFINITION

The International Classification of Diseases Revision 10 (ICD-10) defines a maternal death as follows. "A maternal death is defined as the death of a woman while pregnant or within 42 days of termination of pregnancy, irrespective of the duration and site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management but not from accidental or incidental causes." Measuring maternal deaths thus involves the determination of cause of death, an issue not addressed elsewhere in this manual. Maternal deaths are divided into direct obstetric deaths (of which the major specific causes are haemorrhage, obstructed labour, eclampsia, sepsis and consequences of abortion) and indirect obstetric deaths (pregnancy-related deaths among women with a pre-existing or newly developed health problem exacerbated by the pregnancy or delivery).

The measurement of maternal mortality represents a major problem for countries lacking largely complete birth and death registration (Graham, Ahmed, Stanton *et al*. 2008) not only because deaths are not recorded but also because of the need to ascertain cause of death (see for example Mathers, Fat, Inoue *et al*. (2005)). Cause of death is best determined by a physician present close to the time of death, but many deaths occur without the presence of a doctor. Further, even when a doctor does certify the death, deaths that occur outside of the labour ward may be incorrectly ascribed to a non-maternal cause. Some progress has been made in recent years with the development and application of verbal autopsy methods, whereby family members are asked to report signs and symptoms surrounding the death, but there is still considerable controversy about how well such methods work (Chandramohan, Rodrigues, Maude *et al*. 1998). The description of verbal autopsy instruments and analysis is beyond the scope of this manual.

In part because of the difficulty of identifying true maternal deaths, ICD-10 also defines a pregnancy-related death as one that occurs during pregnancy, delivery or the 42 days after the end of the pregnancy, regardless of cause of death. The category pregnancy-related death thus includes all maternal deaths plus the accidental or incidental deaths excluded from the category "maternal". The advantage of the pregnancy-related category is that it appears to be easier to implement; it only requires information on the timing of death relative to a pregnancy, without specific knowledge of true cause of death. The disadvantages are that measures of pregnancy-related mortality are frequently misinterpreted as measures of maternal mortality, and that any trend in pregnancy-related mortality not due to maternal causes

will limit the value of the measure for tracking impact of maternal health interventions. Demographic surveys generally measure pregnancy-related deaths and avoid the necessity for cause ascertainment.

There is active debate, and no consensus, as to the relationship that exists in practice between a reported number of pregnancy-related deaths and some unknown true number of maternal deaths. Clearly by definition the true number of pregnancy-related deaths has to be the same as, or larger than, the true number of maternal deaths, since all maternal deaths are pregnancy-related, but not all pregnancy-related deaths are maternal. In practice, however, the situation is less clear-cut because of possible reporting errors. There are those who argue that reported pregnancy-related deaths exceed true maternal deaths (Garenne, McCaa and Nacro 2008; Stecklov 1995) whereas others argue that pregnancy-related deaths are likely to be under-reported because, for example, a respondent may not have known that the deceased was pregnant at the time of death (Shahidullah 1995; Wilmoth 2009). The relationship could, therefore, go either way. This manual does not come down firmly in support of either of these views, but instead emphasizes that good practice requires that measures be labelled correctly. Thus a measure based on reported pregnancy-related deaths should be reported as a measure of pregnancy-related mortality, whereas a measure based on what are thought to be true maternal deaths (identified by a verbal autopsy for example) should be reported as a measure of maternal mortality.

## BASIC MEASURES OF MATERNAL MORTALITY

There are two common measures of maternal mortality (and corresponding measures for pregnancy-related mortality). They are the Maternal Mortality Ratio (MMR), the target for MDG-5, and the Maternal Mortality Rate (here abbreviated as MMRate). There are two other measures that will often be encountered: the proportion of deaths of women of reproductive age that are maternal (often abbreviated as PMDF), and the lifetime risk of dying a maternal death (LTR). The latter measure is used primarily for advocacy purposes.

### Maternal Mortality Ratio

The MMR is the number of maternal deaths in a period per 100,000 live births in the same period. Note the use of live births rather than pregnancies in the denominator. The MMR is primarily a measure of obstetric risk, roughly the risk of dying per 100,000 risky events.

### Maternal Mortality Rate

The MMRate is a cause-specific mortality rate. It is the number of maternal deaths in a period per 1,000 person-years lived by the female population of reproductive age (usually ages 15–49).

The MMR and MMRate share a numerator, and have a simple relationship to one another:

$$\begin{aligned} MMR &= \frac{MD}{LB} \times 100{,}000 \\ &= \frac{MD}{FPRA} \times 1{,}000 \times \frac{FPRA}{LB} \times 100 \\ &= MMRate \times \frac{100{,}000}{GFR} \end{aligned}$$

where, for a given time period, *MD* is maternal deaths, *LB* is live births, *FPRA* is the person-years lived by the female population of reproductive age, and *GFR* is the General Fertility Rate expressed per 1,000 women of reproductive age.

### Proportion Maternal of Deaths of Women of Reproductive Age

The *PMDF* is *MD/FDRA*, where *FDRA* is the number of deaths of women of reproductive age. It is used primarily in modelling exercises (for example, Hill, Thomas, AbouZahr *et al.* (2007), Hogan, Foreman, Naghavi *et al.* (2010), Wilmoth, Zureick, Mizoguchi *et al.* (2010) and Wilmoth, Mizoguchi, Oestergaard *et al.* (2012)) but is also of some value for data quality assessment (see below).

### Lifetime Risk

The LTR is usually implemented as the risk of dying from a maternal cause from age 15 onwards. Wilmoth (2009) suggests calculating the measure per 1,000 women reaching age 15; i.e., as

$$LTR = \frac{(T_{15} - T_{50})}{l_{15}} \times MMRate$$

where $T_{15}$ and $T_{50}$ are the person-years lived above ages 15 and 50 respectively, and $l_{15}$ is the survivors to age 15, in an appropriate life table for the population in question.

Each of the four measures above has a pregnancy-related corollary, calculated by replacing maternal deaths by pregnancy-related deaths.

## DATA SOURCES

Other than civil registration, there are two widely used approaches to the collection of data needed to measure pregnancy-related mortality: the full sibling history (FSH); and a large household survey or census that collects data on recent household deaths (HSHD). The summary sibling history (Graham, Brass and Snow 1989) is now rarely used, partly because it produces estimates that represent averages over very long time frames.

### The full sibling history

A full sibling history (FSH) involves complex and detailed data collection, requiring very careful training and supervision of field staff to be executed correctly. It is therefore not an appropriate methodology to include in a census. The FSH has been widely included as the "Maternal Mortality Module" in DHSs since 1991, and has also been included in some other household surveys. The FSH collects information from eligible respondents. In most DHSs, women eligible for the birth history are also those eligible for the FSH, but some surveys have also collected FSHs from eligible male respondents. Information is collected about all brothers and sisters born to the same mother. The FSH can thus be thought of as the respondent's mother's full birth history, excluding the respondent herself (or himself). In the DHS, the information collected about each sibling is: name; sex; whether still alive; if still alive, age in years; if dead, how many years ago did the sibling die and how old was he or she at death. For deaths of women of reproductive age, additional questions enquire whether the sister died (i) while pregnant; (ii) during childbirth; or (iii) within 42 days or 2 months of the end of a pregnancy.

It will be clear from the above that measures calculated from an FSH are of pregnancy-related mortality, not maternal mortality. The sibling history does not lend itself to the application of a verbal autopsy (which would be necessary for calculating maternal mortality), because a sister of reproductive age reported to have died may well have died in a different household than the respondent, who thus may have little direct knowledge of signs and symptoms preceding the death. It is generally not feasible to try to identify the household where the death occurred and conduct a verbal autopsy with a member of that household.

The FSH provides information on pregnancy-related deaths and female exposure, and thus a basis for estimating pregnancy-related mortality rates. If pregnancy-related mortality ratios (PRMRs) are to be calculated, information must also be available on live births. A typical DHS collects a full birth history (FBH) as well as an FSH, so this is usually not a problem.

### The summary sibling history

The use of information on sibling survival to estimate maternal mortality was first proposed by Graham, Brass and Snow (1989). They proposed using a summary sibling history. Such a summary history collects information by sex on the aggregate number of siblings the respondent had, the number who survived to age 15 (or first marriage), and – for sisters who died after age 15 – whether they were pregnant, in childbirth, or in the 42 days post-partum when they died. This method is not recommended for use. The sisters of a respondent can differ in age from the respondent herself by plus or minus 30 years, with the result that the deaths of sisters can be spread over a very long time period prior to a survey. Reference dates of maternal mortality estimates derived from summary sibling histories are thus located well in the past (on average as much as 12 years before the survey), making them of limited practical value. As a consequence the method will not be described further.

### The census or large household survey collecting data on recent household deaths

Censuses in the 1970s and 1980s in countries lacking complete civil registration often collected information on recent household deaths, usually those that occurred in the last 12 months. Concerns about data quality limited the use of such questions in the 1990 and 2000 rounds of censuses, but renewed interest in adult mortality and specifically in maternal mortality led to a sharp increase in their use in the 2010 round. A common format for such questions is to ask whether any usual household member died in the preceding 12 months (sometimes the question refers to a different period such as the time since a fixed date or memorable event). If the answer is yes, the deceased's name, sex and age at death are recorded. If the death is of a woman of reproductive age, an additional question or questions about the timing of death relative to pregnancy are asked, namely did the deceased die

while pregnant, during delivery, or in the 6 weeks (sometimes 2 months) after the end of the pregnancy? The methodology is reviewed by Stanton, Hobcraft, Hill *et al*. (2001), and experience with it is reviewed by Hill, Queiroz, Stanton *et al*. (2007) and Hill, Queiroz, Wong *et al*. (2009).

As generally used, these questions on recent deaths identify pregnancy-related deaths. However, some attempts have been made to follow up reported deaths of women of reproductive age (or a sample of such deaths) with a verbal autopsy to identify true maternal deaths. This has been done in a number of very large household sample surveys (e.g. in the Bangladesh Maternal Morbidity and Mortality Survey 2001 (Hill, El-Arifeen, Koenig *et al*. 2006), but also following at least two censuses (1986 in Iran and 2007 in Mozambique)).

A census or large household survey that collects information on recent household deaths will always record a household roster by age and sex. This roster provides information on denominators for pregnancy-related mortality rates (PRMRates). Additional information on recent fertility will be needed to calculate PRMRatios. This will usually be collected in the form of a question for women of reproductive age as to whether they had a live birth in the year before the survey or a question on the date of each woman's most recent live birth. Information on life-time fertility should also be collected to permit the evaluation and possible adjustment of the data on fertility (see Chapter 7 on fertility estimation using a relational Gompertz model).

**Data collected at health facilities**

A major expense of household surveys is the cost of getting an interviewer to the (correct) household. Much of this expense can be eliminated by taking advantage of respondents coming to the interviewer, such as to give birth in a health facility. Health facilities are also likely to record births and deaths and cause of death that occur at the facility as part of a routine health management information system. However, the problem with such data is selection bias: we can never be sure that the women who give birth in a facility are representative of all mothers. To improve coverage, experiments are being conducted to find out whether health extension workers or the equivalent working in communities can collect adequate data on births and deaths. Such an approach is akin to a sample registration system.

An analysis strategy for facility data has been proposed but not implemented. Starting with the assumption that women who deliver in a health facility (or visit a health facility for some other pregnancy-related condition) are a biased sample of all mothers (it is not clear which way the bias will run, whether women having fewer pregnancy-related health issues or those having more will predominate), if one could estimate the selection probabilities correctly, the statistics collected could be adjusted for bias. For example, women on visiting a health facility could be asked their age, children ever born and children still alive, plus a number of additional questions about their socio-economic condition. The children ever born and children still alive could then be modelled onto the socio-economic structure of the whole population, available for example from a population census. To our knowledge, this approach has never been tested.

## REFERENCES


Chandramohan D, LC Rodrigues, GH Maude and RJ Hayes. 1998. "The validity of verbal autopsies for assessing the causes of institutional maternal death", *Studies in Family Planning* **29**(4):414–422. doi: 10.2307/172253

Garenne M, R McCaa and K Nacro. 2008. "Maternal mortality in South Africa in 2001: From demographic census to epidemiological investigation", *Population Health Metrics* **6**:4. doi: 10.1186/1478-7954-6-4

Graham W, S Ahmed, C Stanton, C Abou-Zahr and O Campbell. 2008. "Measuring maternal mortality: An overview of opportunities and options for developing countries", *BMC Medicine* **6**:12. doi: 10.1186/1741-7015-6-12

Graham W, W Brass and RW Snow. 1989. "Estimating maternal mortality: The sisterhood method", *Studies in Family Planning* **20**(3):125–135. doi: 10.2307/1966567

Hill K, S El-Arifeen, M Koenig, A Al-Sabir, K Jamil and H Raggers. 2006. "How should we measure maternal mortality in the developing world? A comparison of household deaths and sibling history approaches", *Bulletin of the World Health Organization* **84**(3):173–180. doi: 10.2471/BLT.05.027714

Hill K, BL Queiroz, C Stanton and C AbouZahr. 2007. "Measuring maternal mortality via the population census: Experience from Africa," Paper presented at 5th African Population Conference. Arusha, Tanzania, 10–14 December 2007.

Hill K, BL Queiroz, L Wong, J Plata *et al.* 2009. "Estimating pregnancy-related mortality from census data: Experience in Latin America", *Bulletin of the World Health Organization* **87**(4):288–295. doi: 10.2471/BLT.08.052233

Hill K, K Thomas, C AbouZahr, N Walker *et al.* 2007. "Estimates of maternal mortality worldwide between 1990 and 2005: An assessment of available data", *The Lancet* **370**(9595):1311–1319. doi: 10.1016/S0140-6736(07)61572-4

Hogan MC, KJ Foreman, M Naghavi, SY Ahn *et al.* 2010. "Maternal mortality for 181 countries, 1980–2008: A systematic analysis of progress towards Millennium Development Goal 5", *The Lancet* **375**(9726):1609–1623. doi: 10.1016/S0140-6736(10)60518-1

Mathers CD, DM Fat, M Inoue, C Rao and AD Lopez. 2005. "Counting the dead and what they died from: An assessment of the global status of cause of death data", *Bulletin of the World Health Organization* **83**(3):171–177.

Shahidullah M. 1995. "The sisterhood method of estimating maternal mortality: The Matlab experience", *Studies in Family Planning* **26**(2):101–106. doi: 10.2307/2137935

Stanton C, J Hobcraft, K Hill, N Kodjogbé *et al.* 2001. "Every death counts: Measurement of maternal mortality via a census", *Bulletin of the World Health Organization* **79**(7):657–664.

Stecklov G. 1995. "Maternal mortality estimation: Separating pregnancy-related and non-pregnancy-related risks", *Studies in Family Planning* **26**(1):33–38. doi: 10.2307/2138049

Wilmoth J, S Zureick, N Mizoguchi, M Inoue and M Oestergaard. 2010. *Levels and Trends of Maternal Mortality in the World: The Development of New Estimates by the United Nations*. Geneva: WHO, UNICEF, UNFPA and the World Bank. http://www.who.int/reproductivehealth/publications/monitoring/MMR_technical_report.pdf

Wilmoth JR. 2009. "The lifetime risk of maternal mortality: Concept and measurement", *Bulletin of the World Health Organization* **87**(4):256–262. doi: 10.2471/BLT.07.048280

Wilmoth JR, N Mizoguchi, MZ Oestergaard, L Say *et al.* 2012. "A new method for deriving global estimates of maternal mortality", *Statistics, Politics, and Policy* **3**(2):Article 3. doi: 10.1515/2151-7509.1038

# Chapter 29 Estimation of pregnancy-related mortality from survival of siblings

*Kenneth Hill*

## DESCRIPTION OF METHOD

The use of full sibling history data to estimate overall adult mortality, and the evaluation of such data, is described elsewhere. A full sibling history is analogous to a full birth history: a respondent (usually a female of reproductive age) is asked about each of her siblings born of the same mother. For surviving siblings, sex and age in completed years are recorded; for dead siblings, sex, age at death in completed years and calendar year of death are recorded. The resulting history is, with the addition of the respondent herself, the full birth history of the mother. As with a full birth history, the sibling history allows events (deaths) and exposure time to be arranged in calendar time, and hence the calculation of age-period mortality rates. For pregnancy-related mortality, further information is collected concerning deaths of sisters of reproductive age as to whether the sister died during pregnancy, during delivery, or in the 42 days (or sometimes 2 months) post-partum. This chapter focuses on analysing information on sisters of reproductive age.

One important issue discussed under full sibling histories will be touched on here. The DHS full sibling history asks respondents for the full birth history of their natural mother, excluding themselves. As a result, there is the potential for multiple responses about the same individual. For example, if two daughters of the same mother are interviewed in the same household, there will be multiple reports about other members of the sibship. The DHS bases events and exposure time entirely on reported siblings, not taking into account the exposure time of the (surviving) respondent herself. Further, the events and exposure time of siblings are weighted only by the respondent's sample weight, not taking into account numbers of surviving potential respondents in the sibship. Trussell and Rodriguez (1990) show that if there is no correlation between mortality risks within sibships, this calculation gives an unbiased estimate of overall mortality. Gakidou and King (2006) argue that sibships should include the surviving respondent and should in addition be weighted by the likelihood that they will be reported – that is, by the inverse of the number of potential respondents in the sibship. They also argue that an adjustment should be made for sibships that go unreported because no member remains alive. In a multi-survey analysis of DHS FSH's, Obermeyer, Rajaratnam, Park *et al*. (2010) estimate that the effect of not adjusting for likelihood of reporting can bias overall mortality estimates downwards by 20 percent or so. Masquelier (2012) however argues that the Obermeyer, Rajaratnam, Park *et al*. analysis took into account all surviving siblings, not just potential respondents, and as a result exaggerated the size of any bias. Masquelier recommends using the DHS calculation approach, for reasons outlined below (see also Chapter 27).

## DATA REQUIREMENTS AND ASSUMPTIONS

### Important assumptions

- No correlation exists between mortality risks of women and size of sibship.
- There are no selection effects resulting from migration.

### Tabulations of data required

- Number of women, by five-year age group, from household questionnaire.
- Number of sister deaths by time period (typically 0–6 years) before the survey by five-year age group and by whether during pregnancy, delivery or 42 days/2 months post-partum.
- Number of sister-years of exposure by time period by five-year age group of respondent.
- Age-specific fertility rates and General Fertility Rate for the time period.

We will assume here that the DHS approach is followed. The extraction of summary data from the DHS is beyond the scope of this manual. A standard program exists in the free software CSPro to produce this tabulation from the basic data. A decision that has to be made at the outset of the analysis is the time frame to be used. Most DHSs create the basic table for events and exposure in the seven (0 to 6) years prior to the survey, but sometimes the period used is five years, and sometimes ten years. One consideration in choosing one period rather than another is sampling errors; in a small sample with quite low adult mortality, the period may have to be ten years to avoid very large sampling uncertainty, whereas with a large sample and higher mortality, the period may be reduced to five years. This aspect of the method is discussed further in the section on interpretation of results.

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Data quality assessment for a full sibling history and for recent age-specific fertility rates are described elsewhere. The only data quality assessment specific to the estimation of Pregnancy-Related Mortality is of the information on the proportion of deaths of women of reproductive age that are pregnancy-related (PPRD), and the proportions of pregnancy-related deaths that occur during pregnancy, during delivery and in the 42 days (or two months) post-partum.

No formal methods exist for carrying out such assessments. However, the age pattern of the proportion of deaths that are pregnancy-related should resemble the age distribution of age-specific fertility, since it is births that are the risky events. Risks of pregnancy-related death are generally thought to be rather higher at the extremes of the reproductive age period, so the tails of the proportions pregnancy-related may be somewhat "fatter" than for age-specific fertility.

## CAVEATS AND WARNINGS

It is widely believed that sibling histories tend to under-report mortality, particularly deaths further in the past. One should thus not attempt to interpret trends over time in pregnancy-related mortality from a single data set. Such attempts across data sets would also need to take into account the wide confidence intervals surrounding estimates even for a seven-year window.

## APPLICATION OF METHOD

### *Step 1: Extract tabulations of the data*

As mentioned earlier, software is readily available to extract the tabulations required relating to reported deaths of sisters, their exposure and pregnancy-related deaths. In addition, estimates of age-specific fertility are required. (If the data on sisters are extracted from a DHS, the approach to estimating fertility rates directly from the data is described elsewhere in this manual). Finally, an estimate of the female population by age group enumerated in the household survey is required. We define the following terms:

- ${}_5D_x^s$ – the reported number of sisters reported dying between ages $x$ and $x+5$
- ${}_5PY_x^s$ – the number of person-years lived by sisters between ages $x$ and $x+5$
- ${}_5PRD_x^s$ – the number of pregnancy-related deaths of sisters between ages $x$ and $x+5$
- ${}_5f_x$ – age-specific fertility rates of women aged $x$ to $x+5$
- ${}_5N_x^f$ – the population of women aged between $x$ and $x+5$ as enumerated in the survey

Tabulations of each of the above five quantities are required to implement the method.

### *Step 2: Derive measures of mortality*

The age-specific mortality rate is derived by dividing the reported deaths of sisters by the number of person years lived by those sisters in each age group,

$$ {}_5M_x = \frac{{}_5D_x^s}{{}_5PY_x^s}. \qquad (1) $$

The age-specific pregnancy-related mortality rate in each age group is given by

$$ {}_5PRMRate_x = \frac{{}_5PRD_x^s}{{}_5PY_x^s}. \qquad (2) $$

The proportion of deaths that are pregnancy-related is

$$ {}_5PPRD_x = \frac{{}_5PRD_x^s}{{}_5D_x^s} \qquad (3) $$

and the age-specific pregnancy-related mortality ratio is

$$ {}_5PRMRatio_x = 100{,}000.\left(\frac{{}_5PRD_x^s}{{}_5PY_x^s \cdot {}_5f_x}\right). \qquad (4) $$

***Step 3: Estimate population-level measures***

Estimates of the pregnancy-related mortality rate, and the proportion of deaths that are pregnancy-related in the population of women of reproductive age (taken here as those aged 15–49) are derived by weighting the age-specific rates derived above by the number of women aged 15–49 in the households surveyed. Thus

$$ {}_{35}PRMRate_{15} = \frac{\sum_{x=15,5}^{45} {}_5PRMRate_x \cdot {}_5N_x^f}{\sum_{x=15,5}^{45} {}_5N_x^f} \quad (5) $$

and

$$ {}_{35}PPRD_{15} = \frac{\sum_{x=15,5}^{45} {}_5PPRD_x \cdot {}_5N_x^f}{\sum_{x=15,5}^{45} {}_5N_x^f} \quad (6) $$

and

$$ {}_{35}PRMRatio_{15} = 100{,}000. \frac{\sum_{x=15,5}^{45} {}_5PRMRate_x \cdot {}_5N_x^f}{\sum_{x=15,5}^{45} {}_5f_x \cdot {}_5N_x^f}. \quad (7) $$

## WORKED EXAMPLE

The application of the method is illustrated using data collected in the 2004 Malawi DHS. Women were asked about the survival, and – if dead – timing of death relative to pregnancy of sisters.

***Step 1: Extract tabulations of the data***

Tabulations of the required input data are presented in Table 29.1. The tabulations of reports of sisters' deaths and exposure are for the seven years before the survey.

***Step 2: Derive measures of mortality***

Application of the method is shown in Table 29.2, using the data presented in Table 29.1. The first column of Table 29.2 shows age-specific mortality rates, calculated by dividing female deaths by age (col. i of Table 29.1) by sisters' exposure in years (col. ii of Table 29.1) as described by Equation 1. The second column shows age-specific pregnancy-related mortality rates, calculated as for the all-cause age-specific rates but using pregnancy-related deaths only (col. iii of Table 29.1) in the numerator (Equation 2). Column 3 shows the age-specific proportions of female deaths that were reported to be pregnancy-related (col. iii of Table 29.1 divided by col. i of Table 29.1 – Equation 3). Column 4 shows age-specific pregnancy-related mortality ratios, calculated by dividing pregnancy-related mortality rates (col. ii) by age-specific fertility rates (col. iv of Table 29.1) and multiplying by 100,000 (Equation 4).

It is important to note that the entries in the Total row in Table 29.2 are not calculated by summing events and exposure across age groups. The reason for this is that the age pattern of sister exposure is not the same as the age pattern of the female population of reproductive age. To obtain valid population-level estimates of totals, it is necessary to re-weight the age-specific rates and ratios in Table 29.2 by the proportional female age distribution in col. v of Table 29.1,

| Age group | Sister deaths | Sister exposure | Pregnancy-related deaths | Age-specific fertility | Household female population |
|---|---|---|---|---|---|
| | (i) | (ii) | (iii) | (iv) | (v) |
| 15–19 | 117 | 27,622 | 8 | 0.162 | 2,570 |
| 20–24 | 227 | 29,331 | 52 | 0.293 | 3,036 |
| 25–29 | 299 | 23,763 | 64 | 0.254 | 2,247 |
| 30–34 | 245 | 17,228 | 50 | 0.222 | 1,516 |
| 35–39 | 230 | 12,206 | 33 | 0.163 | 1,122 |
| 40–44 | 177 | 7,892 | 22 | 0.080 | 970 |
| 45–49 | 82 | 4,574 | 12 | 0.035 | 743 |
| **TOTAL** | **1376** | **122,616** | **240** | **0.204*** | **12,204** |

*General Fertility Rate calculated as age-specific fertility rates weighted by age distribution of the female household population

Source: Malawi DHS 2004 Tables 13.2 (p.245) and 13.3 (p.247)

*TABLE 29.1 Input data used to estimate pregnancy-related mortality, Malawi, 2004 DHS*

as described by Equations 5, 6 and 7. The denominator in Equation 7 is also not the General Fertility Rate as normally calculated (births divided by female population aged 15 to 49) but rather the age distribution-weighted sum of age-specific fertility rates.

***Step 3: Estimate population-level measures***

Table 29.3 compares the proportion of births in each 5-year age group (calculated as the product of the household female population and the age-specific fertility) to the proportion of PRD (calculated as the product of the household female population and the age-specific PRMRate) in each age group.

## DIAGNOSTICS, ANALYSIS AND INTERPRETATION

### Checks and validation

For checking and validating overall estimates of female mortality, see the Chapters 23 and 27 on the analysis of sibling histories. Checking and validating the extra information provided about pregnancy-related mortality depends on plausibility checks. Is the overall proportion of deaths of sisters of reproductive age reported as pregnancy-related plausible, given the estimated pregnancy-related mortality ratio? Is the distribution of pregnancy-related deaths by age plausible given the age pattern of births (the risky events)?

No generally accepted way exists to assess the plausibility of the overall proportion of deaths pregnancy-related. In general, there is a positive association between the

*TABLE 29.2 Adult female mortality rates and pregnancy-related mortality rates, Malawi, 2004 DHS*

| Age group | Age-specific Mortality Rate | Age-specific Pregnancy-Related Mortality Rate | Proportion of Deaths Pregnancy-Related | Age-specific Pregnancy-Related Mortality Ratio |
|---|---|---|---|---|
| | (Column references as from Table 29.1) | | | |
| | = 1000*(i)/(ii) | = 1000*(iii)/(ii) | = (iii)/(i) | = 100000* iii /(ii*iv) |
| 15–19 | 4.24 | 0.29 | 0.0684 | 178.8 |
| 20–24 | 7.74 | 1.77 | 0.2291 | 605.1 |
| 25–29 | 12.58 | 2.69 | 0.2140 | 1060.3 |
| 30–34 | 14.22 | 2.90 | 0.2041 | 1307.3 |
| 35–39 | 18.84 | 2.70 | 0.1435 | 1658.6 |
| 40–44 | 22.43 | 2.79 | 0.1243 | 3484.5 |
| 45–49 | 17.93 | 2.62 | 0.1463 | 7495.8 |
| **TOTAL*** | **11.51** | **1.99** | **0.1681** | **970.7** |

*Note: all the totals in this table are standardized onto the age distribution of the female household population (see text).

*TABLE 29.3 Comparison of proportions of births and pregnancy-related deaths by age group, Malawi, DHS 2004*

| Age | Household female population | Age-specific fertility | Pregnancy-related deaths* | Births | Proportion of: | |
|---|---|---|---|---|---|---|
| | | | | | Pregnancy-related deaths | Births |
| 15–19 | 2,570 | 162 | 0.74 | 416 | 0.031 | 0.167 |
| 20–24 | 3,036 | 293 | 5.38 | 890 | 0.222 | 0.356 |
| 25–29 | 2,247 | 254 | 6.05 | 571 | 0.249 | 0.228 |
| 30–34 | 1,516 | 222 | 4.40 | 337 | 0.181 | 0.135 |
| 35–39 | 1,122 | 163 | 3.03 | 183 | 0.125 | 0.073 |
| 40–44 | 970 | 80 | 2.70 | 78 | 0.111 | 0.031 |
| 45–49 | 743 | 35 | 1.95 | 26 | 0.080 | 0.010 |
| **TOTAL** | **12,204** | | **24.27** | **2,500** | **1.000** | **1.000** |

*Calculated as the product of col(i) of this table and col(ii), divided by 1000, of Table 29.2

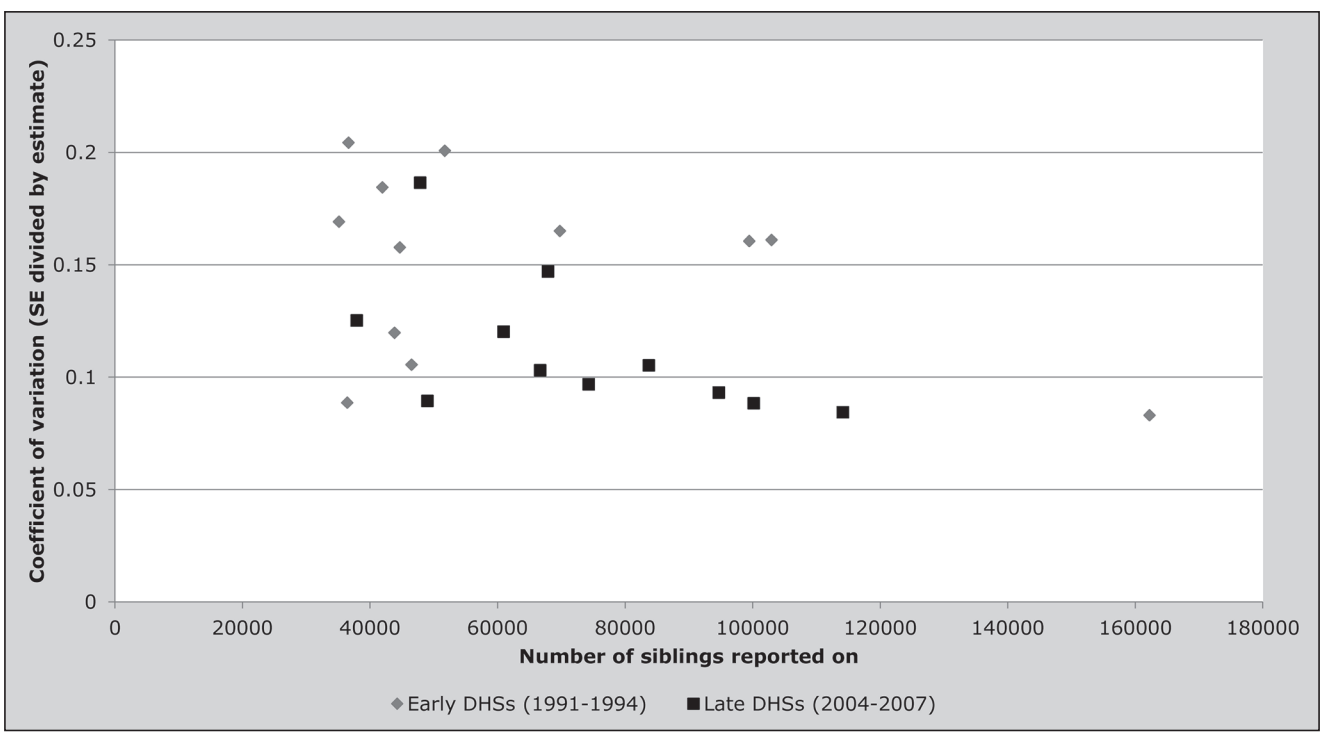
proportion of deaths pregnancy-related and the PRMR, but the association hinges on the level of non-pregnancy-related mortality and provides no useful basis for evaluation. The plausibility of the age distribution of pregnancy-related deaths is assessed in comparison to the distribution of births by age, as shown in Table 29.3. In the case of the 2004 Malawi DHS, the proportions of pregnancy-related deaths contributed by the age groups 15–19 and 20–24 are much lower than the corresponding proportions of births, and the reverse is true over the age of 35. The latter can be plausibly explained by the increasing riskiness of pregnancy and childbirth for women over age 35, but no plausible explanation exists for the large differences under age 25. The suspicion is strong that deaths of sisters under the age of 25 that were actually pregnancy-related have not been reported as such.

**Interpretation**

Interpretation of estimates of pregnancy-related mortality from sibling histories needs to pay close attention to sampling uncertainty and typical data errors. Sampling uncertainty is very large by comparison with DHS estimates of under-5 mortality. Figure 29.1 plots coefficients of variation (standard error divided by the estimate) for DHS estimates of PRMRs by number of siblings reported on. The coefficients of variation are above 0.08 even for very large samples, and above 0.10 for all but a handful of surveys.

Given the large sampling uncertainty, interpretation of sub-national differences or other sub-group differences such as by age is hazardous. Typical data errors, particularly the omission of deaths for time periods more distant from the survey, make any interpretation of trends within data sources questionable. Any conclusions about trends should be based on estimates from two or more surveys for comparable time periods before each survey and taking into account sampling uncertainty.

***FIGURE 29.1*** *Coefficients of variation for PRMRatios by DHS sample size*




Sources: Stanton, Abderrahim and Hill (2000) for estimates prior to the year 2000, and DHS country reports for years 2004 and later

## REFERENCES


Gakidou E and G King. 2006. "Death by survey: estimating adult mortality without selection bias from sibling survival data", *Demography* **43**(3):569–585. doi: 10.1353/dem.2006.0024

Masquelier B. 2013. "Adult mortality from sibling survival data: A reappraisal of selection biases?", *Demography* **50**(1):207–228. doi: 10.1007/s13524-012-0149-1

Obermeyer Z, JK Rajaratnam, CH Park, E Gakidou *et al*. 2010. "Measuring adult mortality using sibling survival: a new analytical method and new results for 44 countries, 1974–2006", *PLoS Medicine* **7**(4):e1000260. doi: 10.1371/journal.pmed.1000260

Stanton C, N Abderrahim and K Hill. 2000. "An assessment of DHS maternal mortality indicators", *Studies in Family Planning* **31**(2):111–123. doi: 10.1111/j.1728-4465.2000.00111.x

Trussell J and G Rodriguez. 1990. "A note on the sisterhood estimator of maternal mortality", *Studies in Family Planning* **21**(6):344–346. doi: 10.2307/1966923

# Chapter 30 Estimation of pregnancy-related mortality from deaths reported by households

*Kenneth Hill*

## DESCRIPTION OF METHOD

If questions are asked in a census or large household survey about deaths in the household in a reference period, and further questions are asked about the timing relative to pregnancy of deaths of women of childbearing age, it is possible to derive estimates of pregnancy-related mortality. If additional information on cause of death is available, for instance from a verbal autopsy enquiry, it is possible to estimate maternal mortality, but this is quite unusual and will not be covered explicitly here.

The use of census or large survey data on recent household deaths to estimate overall adult mortality, and the evaluation of such data, are described elsewhere. Any error in reporting on deaths is likely to have a proportionately similar effect on estimates of pregnancy-related mortality, so evaluation of data quality, and adjustment if needed, are essential parts of the analysis.

## DATA REQUIREMENTS AND ASSUMPTIONS

### Tabulations of data required

- Number of women, by five-year age group from household questionnaire.
- Number of household deaths in the previous 12 months (or similar period) by age and sex.
- For deaths of women of reproductive age (usually 15 to 49), whether the death occurred during pregnancy, delivery or 42 days/2 months post-partum.
- Age-specific fertility rates and General Fertility Rate for the time period.
- If sample or design weights have been provided with the data, they must be applied in the appropriate manner in producing input tables.

### Important assumptions

- That any effect of household dissolution arising from death of a member is trivial.
- (If adjustment of deaths or births is to be made), that errors in reporting deaths and births are proportional by age.
- That any omission of deaths does not vary with whether or not the deaths are pregnancy-related.
- That reporting of deaths as pregnancy-related is approximately accurate.

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Data quality assessment for household deaths and for recent age-specific fertility rates are described elsewhere. The only data quality assessments specific to the estimation of Pregnancy-Related Mortality are of the information on the proportion of deaths of women of reproductive age that are pregnancy-related (PPRD), and the proportions of pregnancy-related deaths that occur during pregnancy, during delivery and in the 42 days (or two months) post-partum.

Potential for data quality assessment for issues other than recording of overall numbers of births and deaths is limited to a plausibility check for reporting of deaths as pregnancy-related. The age pattern of the proportion of deaths that are pregnancy-related should resemble the distribution of births by age of mother.

## CAVEATS AND WARNINGS

It is widely believed that questions on household deaths and on births in the year before a census or survey often under-report true numbers of events. Careful evaluation of coverage of both types of event is essential. In the unlikely event that under-reporting of both types of event is

approximately equal, the pregnancy-related mortality ratio will not be seriously biased, although the pregnancy-related mortality rate will still be biased. However, if data evaluation indicates omission of deaths and births, the data will need to be corrected before a final estimate of pregnancy-related mortality is arrived at.

## APPLICATION OF METHOD

The method is applied in the following stages:

### *Step 1: Extract tabulations of the data*

Instructions as to how to extract the data are outside the scope of this manual. It is usually a simple matter to produce cross-tabulations of the number of women by age group, and the number of deaths of women by age group and timing of death relative to pregnancy (during pregnancy, during delivery, or during the 42 days/2 months post-partum). Where appropriate, the tabulations should be weighted to compensate for under-enumeration (following a post-enumeration survey) and whether the data being analysed are a micro-sample or not. In addition, estimates of age-specific fertility are required. The process of estimating fertility is described elsewhere.

We define the following terms:

- ${}_5D_x^s$ – the reported number of female household members reported dying between ages $x$ and $x+5$
- ${}_5PRD_x^s$ – the number of pregnancy-related deaths of female household members between ages $x$ and $x+5$
- ${}_5f_x$ – age-specific fertility rates of women aged between $x$ and $x+5$
- ${}_5N_x^f$ – the population of women aged between $x$ and $x+5$ as enumerated in the census or large survey

Tabulations of all four of the above variables are required.

### *Step 2: Derive measures of mortality*

The age-specific mortality rate is derived by dividing the reported deaths of women in the household by the number of person years lived by the population in each age group,

$$ {}_5M_x = \frac{{}_5D_x}{{}_5N_x^f}. \quad (1) $$

The age-specific pregnancy-related mortality rate in each age group is given by

$$ {}_5PRMRate_x = \frac{{}_5PRD_x}{{}_5N_x^f} \quad (2) $$

and the age-specific pregnancy-related mortality ratio is

$$ {}_5PRMRatio_x = 100{,}000.\left(\frac{{}_5PRD_x}{{}_5N_x^f \cdot {}_5f_x}\right). \quad (3) $$

The proportion of deaths that are pregnancy-related is

$$ {}_5PPRD_x = \frac{{}_5PRD_x}{{}_5D_x^f}. \quad (4) $$

The proportionate distribution of the PRD over ages 15 to 49 is given by

$$ {}_5DPRD_x = \frac{{}_5PRD_x}{\sum_{x=15,5}^{45} {}_5PRD_x}. \quad (5) $$

## WORKED EXAMPLE

We use as an example the data from the 2008 Malawi Census, specifically the data from questions concerning deaths in the 12 months before the survey/census.

### *Step 1: Extract tabulations of the data*

Table 30.1 is tabulated from individual-level data from a 10 per cent IPUMS sample from the Malawi Census. The table shows the female population of reproductive age by five-year age groups, female deaths reported as occurring in the 12 months before the census, and, for deaths of females aged 15 to 49, whether the death occurred during pregnancy, during delivery, or in the 42 days post-partum.

### *Step 2: Derive measures of mortality*

Application of the method for data available in the form given in Table 30.1 is shown in Table 30.2. All cause age-specific mortality is calculated (col. i) by dividing deaths (col. vi of Table 30.1) by female population (col. i of Table 30.1) as per Equation 1. Note that strictly speaking the deaths pertain to a population on average half a year earlier than that recorded, but the error involved in ignoring this complication is trivial and will be included in an adjustment if the deaths are corrected using one of the appropriate death distribution methods. Age-specific pregnancy-related mortality rates are then calculated by dividing pregnancy-related deaths (col. v of Table 30.1) by female population (col. i of Table 30.1) as in Equation 2. Age-specific pregnancy-related mortality ratios in col. iv of Table 30.2 are then obtained by dividing the age-specific pregnancy-related mortality rates (col. ii) by age-specific fertility rates (col. iii,

obtained from other sources) – as in Equation 3. Age-specific proportions of deaths pregnancy-related are then calculated, dividing pregnancy-related deaths (col. v of Table 30.1) by all-cause deaths (col. vi of Table 30.1) (Equation 4). Finally, the proportional contribution of each age group to overall pregnancy-related deaths is calculated by dividing the number of pregnancy-related deaths in each age group by the total number of pregnancy-related deaths (col. v of Table 30.1) (Equation 5).

## DIAGNOSTICS, ANALYSIS AND INTERPRETATION

### Checks and validation

The key checks for this methodology are the assessment of coverage of adult female deaths and of births (see Death Distribution Methods and Assessment of recent fertility data). The only checks specific to this method are of the distribution of pregnancy-related deaths by age and a very weak check, available in many instances, on the distribution of pregnancy-related deaths by whether they occurred during

*TABLE 30.1 Female population of reproductive age 15–49 and deaths by whether pregnancy-related, Malawi, 2008 Census*

| | | Female deaths | | | | |
|---|---|---|---|---|---|---|
| Age group | Female population | During pregnancy | During delivery | Post-partum | Total pregnancy-related | All deaths |
| | (i) | (ii) | (iii) | (iv) | (v) | (vi) |
| 15–19 | 67,918 | 43 | 25 | 26 | 94 | 235 |
| 20–24 | 69,069 | 68 | 40 | 36 | 144 | 389 |
| 25–29 | 57,478 | 84 | 31 | 32 | 147 | 442 |
| 30–34 | 41,073 | 92 | 24 | 37 | 153 | 471 |
| 35–39 | 29,993 | 56 | 15 | 23 | 94 | 346 |
| 40–44 | 22,294 | 42 | 4 | 14 | 60 | 238 |
| 45–49 | 17,564 | 38 | 3 | 4 | 45 | 185 |
| **TOTAL 15–49** | **310,748** | **423** | **142** | **172** | **737** | **2,306** |

Source: Malawi 2008 Census, 10 per cent sample

*TABLE 30.2 Adult female mortality rates and pregnancy-related mortality rates: Malawi, 2008 Census*

| | Age-specific | | | | | | |
|---|---|---|---|---|---|---|---|
| Age group | Mortality rate | Pregnancy-Related Mortality Rate | Fertility rate | Pregnancy-related Mortality Ratio | Proportion of Deaths Pregnancy-Related | Proportion of Pregnancy-Related Deaths | Proportion of Births in Previous Year |
| | Table 30.1 (vi)/(i) | Table 30.1 (v)/(i) | | 100000* (ii)/(iii) | Table 30.1 (v)/(vi) | Table 30.1 (v)/Sum(v) | |
| 15–19 | 0.00346 | 0.00138 | 0.1108 | 1249.12 | 0.40000 | 0.12754 | 0.14408 |
| 20–24 | 0.00563 | 0.00208 | 0.2464 | 846.13 | 0.37018 | 0.19534 | 0.32584 |
| 25–29 | 0.00769 | 0.00256 | 0.2296 | 1113.89 | 0.33258 | 0.19946 | 0.25267 |
| 30–34 | 0.01147 | 0.00373 | 0.1941 | 1919.15 | 0.32484 | 0.20760 | 0.15264 |
| 35–39 | 0.01154 | 0.00313 | 0.1457 | 2151.04 | 0.27168 | 0.12754 | 0.08367 |
| 40–44 | 0.01068 | 0.00269 | 0.0718 | 3748.34 | 0.25210 | 0.08141 | 0.03065 |
| 45–49 | 0.01053 | 0.00256 | 0.0311 | 8238.13 | 0.24324 | 0.06106 | 0.01046 |
| **Total 15–49** | **0.00755** | **0.00241** | **0.1713** | **1406.89** | **0.33804** | **1.00000** | **1.00000** |

Source: Table 30.1 and Malawi 2008 Census, 10 per cent sample

pregnancy, during delivery, or in the 6 weeks/2 months post-partum. For the assessment of the distribution of pregnancy-related deaths by age, the key comparison is with the distribution of births by age. These two distributions are shown in the last two columns of Table 30.2. The proportions of pregnancy-related deaths through age group 25–29 are lower than the corresponding proportions of births, but above age 30 the reverse is true. This pattern is plausible given accumulating evidence that pregnancy-related mortality risks are broadly similar by age below age 30 but then rise steeply above age 30. It is instructive to compare the patterns with those based on the sibling histories of the 2004 Malawi DHS (described in Table 29.3). In the sibling history example, the proportions of pregnancy-related deaths contributed by mothers under age 25 were much smaller than the corresponding proportions of births. Here, the proportions are smaller, but not so much smaller as to lead to concerns about data accuracy.

It is noted above that the distribution of pregnancy-related deaths by whether they occurred during pregnancy, during delivery, or in the 6 weeks/2 months post-partum is a weak check. It is weak because there is no strong prior as to what this distribution should look like across different settings, and because in practice the distribution is found to vary wildly by data source.

**Interpretation**

The importance of evaluating coverage of adult female deaths and births cannot be over-emphasized. Household deaths in some settings seem to be under-reported by as much as 50 per cent, and such an error would translate into a bias in the pregnancy-related mortality ratio of a similar magnitude. Recent births may also be under-reported, an error that may partially compensate for omission of deaths in the PRMRatio. Death distribution methods suggest that female deaths were under-reported in the 2008 Malawi Census by somewhere between 40 and 50 per cent, whereas application of the relational Gompertz method to the same data indicates births in the year before the census were under-reported by about 18 per cent. The net effect of the two compensating errors would be an under-estimate of the pregnancy-related mortality ratio of about two-fifths.

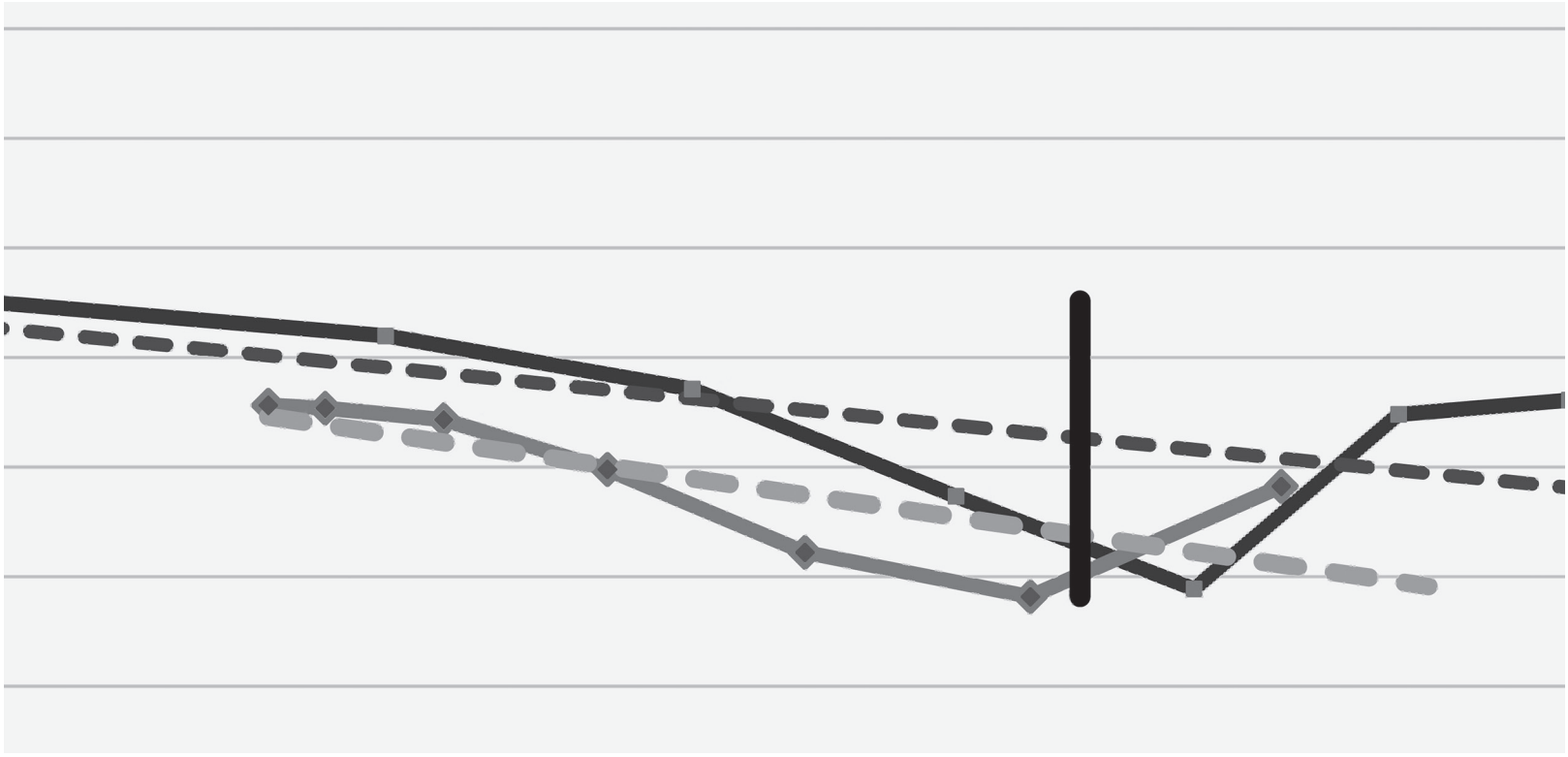
In interpreting information on pregnancy-related mortality for Malawi, it is also essential to remember that Malawi was affected by a substantial HIV epidemic in the late 1990s and early 2000s. Associated HIV-related mortality will affect the results of death distribution methods, as described elsewhere, so the adjustment factor derived above should be used with caution.

One advantage of census data or large census samples is the lack of sampling uncertainty in the results. Thus whereas it is hazardous to draw conclusions from sibling histories about differentials, similar reservations do not apply to the same extent to estimates derived from census data (though even with a census numbers may be small for sub-groups, introducing stochastic error). Also, given the need to evaluate, and often adjust, data from census questions, any estimates will still be subject to considerable uncertainty.

## REFERENCES

Hill K, C Stanton, M Levin *et al*. (2011) *Measuring Maternal Mortality from a Census: Guidelines for Potential Users*. Geneva: World Health Organization.

# Using Models to Derive Life Tables from Incomplete Data



# Chapter 31 Introduction to model life tables

*Tom A Moultrie and Ian M Timæus*

Model life tables are used for comparison in the assessment of empirical estimates of mortality, to smooth or otherwise adjust defective mortality estimates, and to complete the life table when estimates of mortality are available for only a limited range of ages.

The term "to smooth" in this context refers to any procedure for the elimination or minimization of irregularities present in reported data or in preliminary estimates obtained from them. Such smoothing techniques encompass a wide variety of procedures ranging from the fitting of models to simple averaging. Traditional techniques for the smoothing or graduation of age distributions and observed age-specific mortality rates such as the use of cubic splines are well described in the actuarial and demographic literature and are not discussed here. Instead, the focus is on model-based procedures that are suitable for use when the basic data are either defective or incomplete.

In classical demographic analysis, a life table is calculated by converting a complete series of age-specific death rates (${}_nm_x$) into probabilities of dying (${}_nq_x$). From these one can calculate survivorship, $l(x)$, and all the other functions of the life table. In the analysis of census and survey data, however, one often only obtains mortality estimates for part of the age range. For example, mortality estimates made from birth history data (Chapter 15) and sibling history data (Chapter 27) provide no information on the mortality of older children or on adult mortality at age 50 and more. With estimates of this sort, model life tables can be used both to smooth the estimated death rates and to complete the life table by making plausible assumptions about the death rates that prevail at ages at which mortality has not been measured directly.

Furthermore, if one has estimated survivorship indirectly from information on children ever born and surviving (Chapter 16) and on the survival of parents (Chapter 22) or other adult relatives, the results indicate only the level of mortality in each broad age range. In particular, for adults these methods yield conditional survivorship probabilities i.e. probabilities of surviving from age $A$ to age $B$, $l(B)/l(A)$. In this situation, model life tables can be used both to estimate death rates for five-year age groups and to complete the life table by making a plausible assumption about mortality in old age.

The following two chapters describe two approaches to deriving complete life tables from information on mortality in a limited range of ages. Chapter 32 focuses on methods for combining a single estimate of child mortality (${}_5q_0$) and a conditional estimate of adult mortality referring to the same year or period of years to derive a full life table. Several variants of the method are described. Chapter 33 describes a method for combining a series of time-located estimates of child mortality (such as those produced indirectly from data on children ever-born and surviving) and a series of time-located estimates of conditional adult survival (such as those produced by the indirect orphanhood or siblinghood methods) to produce a life table for a specific point in time.

This manual makes extensive use of relational logit model life tables, firstly, for the evaluation and smoothing of series of estimates of child and adult mortality and, secondly, to combine independent estimates of child and adult mortality and produce full life tables. The system and its properties are described in the next section.

## OVERVIEW OF THE RELATIONAL LOGIT SYSTEM OF MODEL LIFE TABLES

Brass and colleagues (Brass 1964, 1971; Brass and Coale 1968) developed a flexible 2-parameter system of model life tables usually referred to as the logit model life table system. Broadly speaking, the first parameter of this system of models, $\alpha$, captures differences in the level of mortality between populations and the second parameter, $\beta$, variation between populations in the relationship between mortality in childhood and adulthood.

The system of models is a relational one. In other words, it is based on a mathematical transformation of the age-specific survivorship function, $l(x)$, which makes it possible to relate two different life tables to each other by means of a simple equation. In particular, Brass discovered that a logit transformation of the probabilities of survival to age $x$, $l(x)$, rendered the relationship between transformed probabilities for different life tables approximately linear.

Thus, if one defines the logit of $l(x)$ as

$$Y(x) = \text{logit}\big(l(x)\big) = -\frac{1}{2}\ln\left(\frac{l(x)}{1-l(x)}\right), \qquad (1)$$

the following linear relationship is approximately true for all ages $x$:

$$Y(x) = \alpha + \beta Y^*(x) \qquad (2)$$

where $Y(x)$ and $Y^*(x)$ are the logits of survivorship by age, $l(x)$ and $l^*(x)$, in two different life tables, and $\alpha$ and $\beta$ are constants.

Those familiar with logistic regression will recognize $Y(x)$ as being half the log odds of dying between birth and age $x$ since

$$\text{logit}\big(l(x)\big)\times 2 = -\ln\left(\frac{l(x)}{1-l(x)}\right) = \ln\left(\frac{1-l(x)}{l(x)}\right)$$
$$= \ln\left(\frac{{}_xq_0}{1-{}_xq_0}\right).$$

If Equation 2 held for any pair of life tables, this would imply that all life tables could be generated from a single baseline or standard life table, $l^s(x)$, using an appropriate pair of values of $\alpha$ and $\beta$. In fact, Equation 2 is only approximately satisfied by pairs of actual life tables, but the approximation is close enough to warrant use of the relationship to study and model observed mortality schedules.

Before describing how to use Equation 2 to generate model life tables, a word about the meaning of the parameters $\alpha$ and $\beta$ is in order. Consider the set of life tables that can be generated starting with some baseline life table $l^s(x)$ and calculating $Y(x)$ for different values of $\alpha$ and $\beta$. If $\beta$ is held constant and equal to 1, changing $\alpha$ will either increase or decrease survivorship at every age. Thus changing $\alpha$ will produce life tables whose shapes are essentially the same as that of the $l^s(x)$ life table used to generate them, but whose overall levels differ. If, on the other hand, $\alpha$ is fixed at 0 and $\beta$ is allowed to vary, the resulting life tables will no longer display the same shape as $l^s(x)$. All of the *derived* tables will intersect at a single point located somewhere in the central portion of the age range, where $l^s(x) = 0.5$ and $Y^s(x) = 0$. Therefore, their probabilities of survival will be either lower at younger ages and higher at older ages or lower at younger ages and higher at the older than the standard survivorship probabilities $l^s(x)$ from which they are generated. Hence, $\beta$ modifies the shape of the generated mortality schedule rather than its level. Simultaneous changes of $\alpha$ and $\beta$ will bring about changes in both the level and shape of the survivorship function being generated.

From Equation 1,

$$l(x) = \frac{1}{1+\exp\big(-2Y(x)\big)}$$

and combining this with Equation 2:

$$l(x) = \frac{1}{1+\exp\Big(-2\big(\alpha+\beta Y^s(x)\big)\Big)}. \qquad (3)$$

Thus, for any series of $l^s(x)$ values defining a standard life table, another series $l(x)$ can be obtained for each pair of $\alpha$ and $\beta$ values. (Note that, at the endpoints of the age range, Equation 3 cannot be used to calculate $l(x)$; $l(0)$ and $l(\omega)$ should be set to 1 and to 0, respectively).

Equation 3 can be used to generate families of model life tables from an appropriate standard life table, $l^s(x)$. Potentially, any life table can be used as a standard. For example, one might use a reliable life table for the population concerned at some other date or a life table from a neighbouring country. When no appropriate or reliable such life table exists, however, a model life table taken from the Princeton regional series (Coale, Demeny and Vaughan 1983), or the *UN Model Life Tables for Developing Countries* (UN Population Division 1982) is frequently used as a standard. The derivation of the standards used in this manual are described in the next section of this chapter.

Because of the mathematical simplicity of Equations 2 and 3, logit model life tables based on any standard can be readily calculated in a spreadsheet, doing away with the need for volumes of published tables. The simple mathematical form of Equation 3 also simplifies the use of relational logit model life tables for simulation purposes and for projecting mortality. If the past and current mortality schedules of a population are known, trends in the $\alpha$ and $\beta$ parameters

can be determined by using the logit model life table system to fit each mortality schedule, and with some caution the trends in these two parameters can be projected to generate estimates of future mortality.

## DESCRIPTION OF THE MODEL LIFE TABLES USED IN THE MANUAL

All the logit life tables used in this manual are based on a common set of standard life tables. These standard life tables are taken from the Princeton regional model life tables (North, South, East and West) and the UN model life tables for developing countries (General, Latin American, Chilean, South Asian, Far Eastern), by sex. They all have an expectation of life at birth of 60 years. The original life tables have been modified, extended and enhanced over time to extend them to older ages. We make use of these updated tables, which were developed by the UN Population Division (2010) and used by them in their population projections. These life tables provide values of $l(x)$ and $L_x$ (amongst other quantities) for ages 0, 1, 5, 10, ..., 130.

For the standards based on the Princeton regional model life tables, values for $l(2)$, $l(3)$ and $l(4)$ were generated by applying the proportionality factors presented by Coale, Demeny and Vaughan (1983: 21) to $l(1)$ and $l(5)$. For the standards based on the UN model life tables for developing countries, deaths between the ages of 1 and 5 were distributed by single years of age in the same proportion as those deaths in the original sex- and region-specific life tables.

Some methods of child and adult mortality estimation require joint-sex life tables (that is, life tables for males and females combined). As these life tables (or their implementation) are not particularly sensitive to the sex ratio at birth, a sex ratio at birth of 105 (boys per 100 girls) is used. Joint-sex life tables were then derived by appropriate weighting of the sex-specific life tables:

$$l^c(x) = \frac{(1.05)l^m(x) + l^f(x)}{2.05}$$

where $l^c(x)$ represents the number of survivors at age $x$ in the joint-sex life table and $l^m(x)$ and $l^f(x)$ are the equivalent life table values for men and women respectively.

As these life tables are used almost exclusively in a relational context (as originally set out by Brass (1971), standard logits of the $l^s(x)$ values were calculated for all ages above zero by means of the formula

$$Y^s(x) = 0.5 \ln\left(\frac{1 - l^s(x)}{l^s(x)}\right).$$

Values of these logits can be downloaded from the *Tools for Demographic Estimation* website.

## CHOOSING AN APPROPRIATE STANDARD

A crucial decision to be made when implementing methods based on model life tables, or when combining estimates from different methods into a single life table based on the relational model life table system is the choice of standard life table to be used in the calculations.

The nine standard life tables (four Princeton regional model life tables; five UN developing country life tables) used in this manual exhibit markedly different mortality patterns. Figure 31.1 shows the relative balance of child and adult mortality for the combined sex standard life tables – all of which have a life expectancy at birth of 60 years. The index of child mortality is ${}_5q_0$, the probability of dying before exact age 5; the index of adult mortality is ${}_{45}q_{15}$, the conditional probability of dying between exact ages 15 and 60.

Thus, for example, the UN Far Eastern table is revealed to have very high adult mortality and very low child mortality relative to the other tables used, while – at the other extreme – the Princeton South and UN South Asian tables have relatively low adult mortality but very high child mortality.

A second important dimension on which the tables differ is in the balance between infant mortality (before the first birthday) and child mortality (between exact ages 1 and 5). Comparing these rates (Figure 31.2), it can be seen that while the UN Far Eastern and UN Chilean tables have roughly equivalent levels of child mortality, the level of infant mortality in the two standards is very different.

Ideally, a standard life table should be selected for any application that describes well the relative balance between infant and child mortality, on the one hand, and between under five mortality and adult mortality on the other. Thus, if there are reasonable estimates of the mortality pattern for a given country, the best standard life table can be selected by comparing the observed pattern to those embodied by the model tables. But in populations on which little or no reliable information on mortality by age is available, the analyst can do little more than guess which pattern would be most appropriate.

In situations where nothing is known about the age pattern of mortality, use of either the Princeton West or the UN General standard is recommended because of the relatively wider data base from which these tables were derived. Moreover, the UN General pattern, in particular, appears to represent something close to an average pattern in terms of characteristics plotted in Figures 31.1 and 31.2.

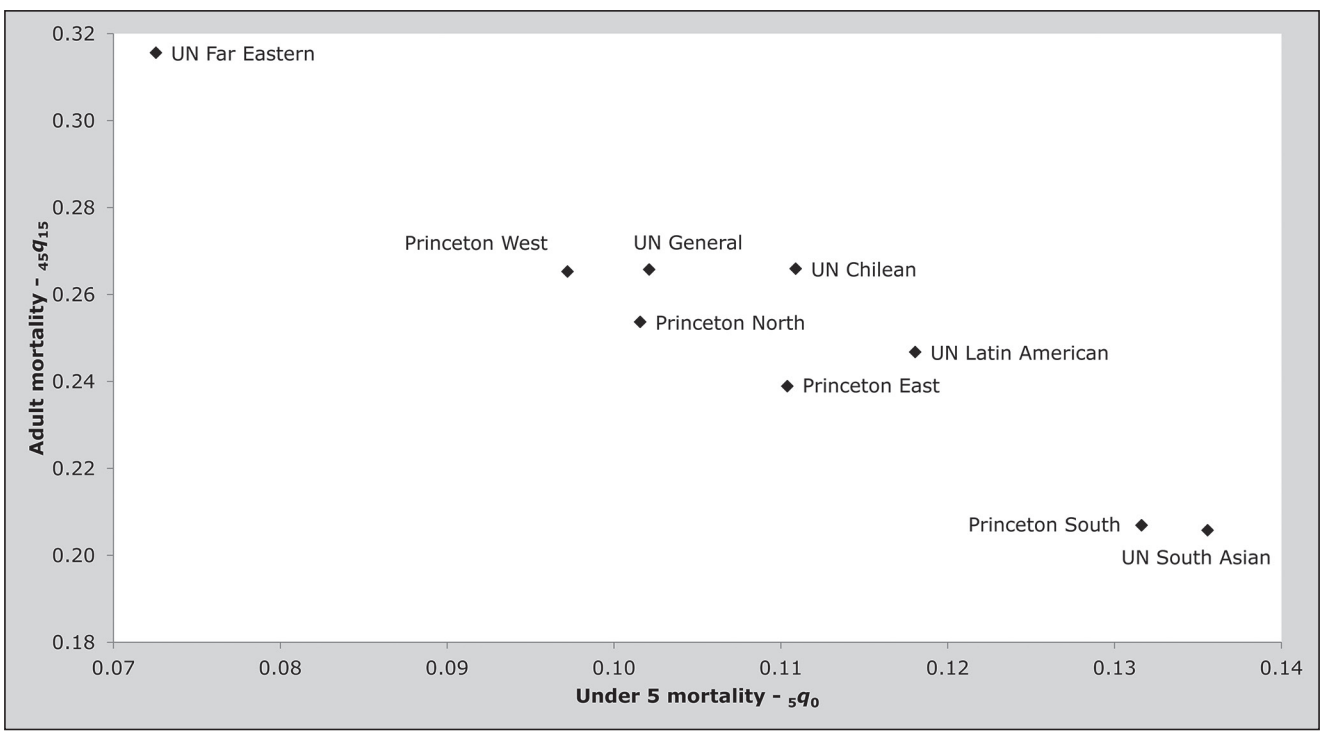
As for the "extra information" that might permit a more enlightened selection, it can vary considerably in type and quality. It might range, for example, from estimates of age-specific mortality rates derived from vital registration data to knowledge of some fairly general facts, such as the prevalence and typical duration of breast-feeding in the population, or an estimate of tuberculosis prevalence.

When a set of observed age-specific mortality rates is available (preferably a set adjusted according to a death distribution method such as those described in Chapters 24 and 25), a model mortality pattern may be chosen by comparing the logits of the observed $l(x)$ values to those in the different standard model life tables. This comparison may be carried out by plotting the observed values of $Y(x)$ against those derived from the different standards, and choosing as the preferred standard that model life table that exhibits the most linear relationship between the two sets of values.

According to the description given above of the mortality patterns contained in the different standard life tables, it is evident that they differ most markedly in their values at early ages and in the relation between infant ($_1q_0$) and child ($_4q_1$) mortality. It follows that quite different child mortality estimates may be obtained from the same information according to which family is selected as representative. Furthermore, in this case, sound external evidence to inform the selection of a standard can be hard to obtain, mainly because infant deaths are very often grossly underreported. In the absence of adequate empirical data for selecting a suitable standard life table, a few general guidelines can be proposed to help narrow the possibilities and lead to a reasonable choice:

a) In a population where breast-feeding is common practice and where weaning occurs at a relatively late age

*FIGURE 31.1 Relationship between adult and under five mortality for different standard life tables with a life expectancy at birth of 60 years*



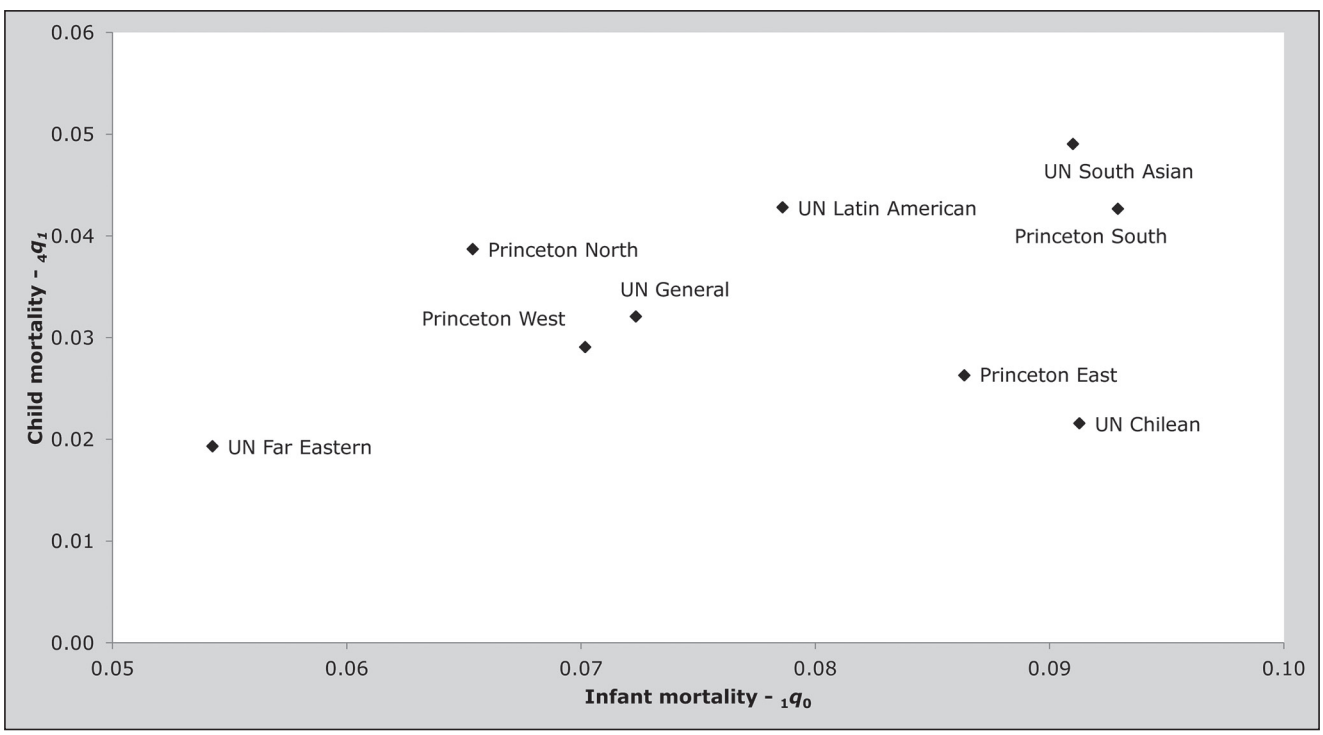
(12 months or over), one may reasonably expect child mortality ($_4q_1$) to be relatively high compared with infant mortality ($_1q_0$) since breast-feeding may successfully prevent deaths due to malnutrition and infectious diseases among young infants. When weaning takes place, however, the child is less protected from these perils and is more likely to die. In these cases, mortality in childhood is likely to be well represented by the Princeton North or UN General standard. Yet, it cannot be inferred from these observations that these tables will provide an appropriate mortality model for other sections of the age range. Only independent information on mortality in adulthood is able to establish this fact.

b) In some populations today, breast-feeding has been abandoned by a high proportion of the female population; and, from a very early age, infants are fed unsterilized and often inadequate rations of "milk formula". When this practice is adopted by women living in relatively unhealthy conditions and exacerbated by poor care at delivery and immediately after birth (perhaps leading to a high incidence of neonatal tetanus), infant mortality can be high relative to mortality later in childhood. In such conditions, the Princeton East or UN Chilean tables may be a good representation of mortality in childhood. The caution in the previous paragraph about whether these tables may adequately describe the balance between adult and child mortality applies.

c) Early weaning may not be the only cause of malnutrition which results in a high infant and child mortality. In some populations, breast-feeding is nearly universal but both levels of hygiene and children's nutritional status are poor and both infant and child mortality are high. For such least developed countries, either the Princeton South or UN South Asian model life tables may be the most appropriate.

d) In the absence of data adequate to determine the most suitable family of model life tables to use for a particular country, one may select the same family as that employed for a neighbouring country with similar cultural and socio-economic characteristics.

***FIGURE 31.2*** *Relationship between infant and child mortality for different standard life tables with a life expectancy at birth of 60 years*



e) For the reasons given earlier, if little is known about the population under study, the Princeton West or UN General standard is recommended.

From these remarks, it is clear that the knowledge about mortality patterns is still fairly limited and that, certainly, better information concerning the mortality experience of populations in developing countries is needed to assess the adequacy of the models now available.

## ALTERNATIVE SYSTEMS OF MODEL LIFE TABLES

Two of the methods of fitting life tables to observed data presented in the next chapter make use of two somewhat different approaches to the modelling of mortality patterns to that pioneered by Brass. These two alternative systems of model life tables are described briefly below. The interested reader is referred back to the source texts for further information.

### The modified logit system

Murray and colleagues (Murray, Ferguson, Lopez *et al.* 2003) proposed a modified system of relational logit model life tables based on a single global standard life table and additional two sets of age-specific coefficients $\gamma(x)$ and $\theta(x)$:

$$Y(x) = \alpha + \beta Y^s(x) + \gamma(x)\left(1 - \left(\frac{l(5)}{l^s(5)}\right)\right) + \theta(x)\left(1 - \left(\frac{l(60)}{l^s(60)}\right)\right).$$

As before, $Y(x)$ denotes the logit transform of $l(x)$, so the first two elements are those of the Brass system of logit model life tables. The first of the two additional sets of coefficients, $\gamma(x)$, adjusts for the level of under-five survivorship relative to the standard, while $\theta(x)$ does the same for the level of adult survivorship relative to the standard.

Despite superficially appearing to be a 4-parameter model, this modification of the logit system of models actually remains a 2-parameter one. Because $\gamma(5)$, $\theta(5)$, $\gamma(60)$ and $\theta(60)$ are all set to zero by definition, $\alpha$ and $\beta$ fully define $l(5)$ and $l(60)$ in the fitted model and, thereby, the two sets of age-specific deviations from the standard pattern of mortality, $\gamma(x)$ and $\theta(x)$. In effect, these deviations serve to reduce the impact on mortality in infancy and old-age of using a value of $\beta$ other than 1 to model the relationship between mortality in childhood and adulthood as a whole.

Users of this system of models should be alert to the fact that the values of $\gamma(x)$ and $\theta(x)$ published in the 2003 paper are reversed with respect to sign. Therefore, the parameters tabulated in Table 3 of that paper should be multiplied by –1 before using them.

### The log-quadratic system

An alternative 2-parameter system of model life tables has been published recently by Wilmoth, Zureick, Canudas-Romo *et al.* (2012). It uses age-specific scalar constants $a(x)$, $b(x)$, $c(x)$ and $v(x)$ and parameters $h$ and $k$ in the following relationship:

$$\ln\left({}_n m_x\right) = a(x) + b(x)h + c(x)h^2 + v(x)k\,.$$

Values of $a(x)$, $b(x)$, $c(x)$ and $v(x)$ were derived from the mortality data contained in the Human Mortality Database, leaving two parameters ($h$ and $k$) with which to fit the model to empirical estimates of mortality.

The first parameter, $h$, measures the overall level of mortality and is defined as the log of the observed ${}_5q_0$. The second parameter $k$ (in combination with $v(x)$) captures the deviation of the observed age pattern of mortality from that of a standard population. In practice, it is chosen to fit one index or a series of observed indices of adult mortality (e.g. ${}_{45}q_{15}$ or ${}_{30}q_{30}$).

## REFERENCES


Brass W. 1964. *Uses of census or survey data for the estimation of vital rates*. Paper prepared for the African Seminar on Vital Statistics, Addis Ababa 14–19 December 1964. Document No. E/CN.14/CAS.4/V57. New York: United Nations.

Brass W. 1971. "On the scale of mortality," in Brass, W (ed). *Biological Aspects of Demography*. London: Taylor and Francis, pp. 69–110.

Brass W and AJ Coale. 1968. "Methods of analysis and estimation," in Brass, W, AJ Coale, P Demeny, DF Heisel*, et al.* (eds). *The Demography of Tropical Africa*. Princeton NJ: Princeton University Press, pp. 88–139.

Coale AJ, P Demeny and B Vaughan. 1983. *Regional Model Life Tables and Stable Populations.* New York: Academic Press.

Murray CJ, BD Ferguson, AD Lopez, M Guillot, J Salomon and O Ahmad. 2003. "Modified logit life table system: principles, empirical validation, and application", *Population Studies* **57**(2):165–182. doi: 10.1080/0032472032000097083

UN Population Division. 1982. *Model Life Tables for Developing Countries.* New York: United Nations, Department of Economic and Social Affairs, ST/ESA/SER.A/77. http://www.un.org/esa/population/techcoop/DemMod/model_lifetabs/model_lifetabs.html

UN Population Division. 2010. *World Population Prospects 2012: Extended Model Life Tables.* New York: United Nations, Department of Economic and Social Affairs. http://esa.un.org/wpp/Model-Life-Tables/data/MLT_UN2010-130_1y.xls

Wilmoth JR, S Zureick, V Canudas-Romo, M Inoue and C Sawyer. 2012. "A flexible two-dimensional mortality model for use in indirect estimation", *Population Studies* **66**(1):1–28. doi: 10.1080/00324728.2011.611411

# Chapter 32 Fitting model life tables to a pair of estimates of childhood and adult mortality

*Ian M Timæus and Tom A Moultrie*

## DESCRIPTION OF THE METHOD

It is often impossible in the analysis of mortality in countries with limited and defective data to derive complete series of age-specific death rates from the available data. Most countries, however, have collected data that can be used to estimate child mortality. In particular, it is usually possible to estimate the under-five mortality rate, ${}_5q_0$. In many countries, moreover, it is possible to use either death registration statistics or census or survey data to estimate adult mortality. Usually, the resulting estimates for adults measure conditional survivorship in adulthood over some broad range of ages (for example, ${}_{45}p_{15}$, the probability of surviving from exact age 15 to exact age 60). This chapter sets out how to fit a relational model life table to pairs of such estimates of child and adult mortality that refer to the same year or period of time.

When one can estimate both child and adult mortality in a population, it is possible to fit a 2-parameter model life table to any pair of such estimates that takes on their observed values. Thus, 2-parameter models make full use of the available data in this situation and, because they reproduce the observed relationship between the level of mortality in childhood and adulthood, are likely to represent the age-specific mortality schedule of the population in question far better than a 1-parameter model fitted to data on mortality in childhood alone.

The approaches set out below can be applied to any two estimates of child and adult mortality, provided that the estimate of child mortality is expressed as the probability of survival from birth. In most situations, however, the observed measure of child mortality will be the under-five mortality rate, ${}_5q_0$. By contrast, the estimate of adult mortality is usually a conditional probability, that is, the probability of survival to some age $x+n$, conditional on having survived to age $x$, i.e. ${}_np_x$. In most cases this base age in adulthood is greater than the upper age limit to which child survival is measured. This makes it impossible to straightforwardly convert the conditional measures of survivorship in adulthood into unconditional ones. Thus, the methods for fitting relational model life tables explained by the introductory descriptions of the models in a number of textbooks cannot be applied and more complicated fitting methods are required.

The chapter describes methods than can be used to fit two different types of 2-parameter model life table to a pair of estimates of childhood and adult mortality. The first set of methods is based on the relational logit system of model life tables, presented in Chapter 31. Three variants to this approach are presented. The 'Splicing method' uses separate 1-parameter logit model life tables to represent child and adult mortality (assuming that $\beta$, the shape parameter in the system, equals 1 in each model) and grafts the model life table for adults onto the childhood model at age 15. Second, the 'Brass logit' method uses both the level ($\alpha$) and shape ($\beta$) parameters of the system of models to determine a life table relative to a chosen standard. Third, the 'Modified logit' method,' proposed by Murray, Ferguson, Lopez *et al.* (2003 again fits a relational logit model life table with parameters $\alpha$ and $\beta$, but uses its own standard life table and further adjusts the estimates to reduce the impact that varying $\beta$ has on infant mortality and mortality in old age.

The second set of 2-parameter model life tables is the 'log-quadratic' system proposed recently by Wilmoth, Zureick, Canudas-Romo *et al.* (2012. This is a regression-based system of models of $\ln({}_5m_x)$. It is not explicitly relational in its formulation, although the model is parameterized using the large corpus of mortality data contained in the Human Mortality Database and one could easily re-express

the system as a relational one using one of the models as a standard.

Because iterative methods are required to fit a model life table to data on conditional survivorship in adulthood, detailed worked examples are not provided in the text. The reader is directed to the associated workbook on the manual's website. The final section of the chapter, however, provides a summary of results of the application of all four methods to a set of estimates of mortality for Kenya in the mid-1980s.

## DATA REQUIREMENTS AND ASSUMPTIONS

The methods described in this chapter require as inputs an estimate of infant and child mortality (${}_5q_0$) and an estimate of adult mortality conditional on survival to adulthood (e.g., ${}_{45}q_{15}$ or ${}_{35}q_{15}$). The two estimates should both refer to the same year or other period of time. In principle, measures of adult mortality conditional on survival to ages in early adulthood other than 15 could also be used.

All the methods described in this chapter draw, either implicitly or explicitly, on a standard mortality schedule, which is then modified to fit the observed values. Two of the three methods based on the relational logit system of model life tables system require the analyst to choose which standard life table to use in the derivation of the fitted model life table. Thus, an important assumption made by these methods is that this selection is appropriate. Chapter 31 offers guidance on how to select a suitable standard life table.

## CAVEATS AND WARNINGS

The four approaches usually produce broadly similar estimates of summary measures such as life expectancy at birth. However, important differences exist between the fitted estimates of age-specific mortality produced by the different approaches. These tend to be largest in populations in which the relationship between adult and child mortality differs greatly from the global average. At present, there is no certainty as to whether one of the approaches described below is generally or universally superior to the others. Further research is required in this regard. However, Murray, Ferguson, Lopez *et al.* (2003) provide evidence that the modified logit system tends to produce better results that the use of the original system of 2-parameter model life tables and Wilmoth, Zureick, Canudas-Romo *et al.* (2012) conclude that the log-quadratic and modified logit models perform equally well.

Since all the methods either implicitly or explicitly draw on the mortality experience of a standard population. The methods should not be used to model age-specific mortality in populations with age-specific mortality schedules that differ radically from that of the selected standard. This warning applies especially to populations severely affected by HIV/AIDS as the observed age pattern of mortality in countries with generalized HIV epidemics differs greatly from those in the Princeton and United Nations model life tables and those in the sets of life tables used to derive the modified logit and log-quadratic 2-parameter systems of models.

### Using *Solver* in Microsoft Excel

The methods described in this chapter are designed to be applied when the only measures of adult mortality available are conditional on survivorship to some base age in early adulthood. With such estimates, one cannot determine analytically what estimates of the model parameters will produce the best fitting model life tables. Iterative trial-and-error methods or optimization routines (such as *Solver* in Microsoft Excel) have to be used to apply all the methods described here other than the 'Splicing method'.

*Solver* is not routinely loaded by standard installations of Microsoft Excel. To enable its use, proceed by selecting "File → Options → Add-ins → Manage Excel Add-ins → Go …" and then ensuring that the "Solver Add-in" is ticked.

The specifications of the *Solver* function, and the conditions and constraints that should be adhered to, have been set up in the workbook associated with the methods presented in this chapter. To run the routine on a given worksheet, select "Data → Solver → Solve". Small changes to the specification (in the "Set objective"; the "By Changing Variable Cells"; and the "Subject to the Constraints" textboxes) are required to solve for first the model for men and then the model for women. The appropriate cells for fitting the model life table for women can be identified readily by examining the cells specified for the men.

## METHOD 1: THE SPLICING METHOD

The Splicing method is the simplest of the four methods. It fits a single life table to the estimates of child and adult mortality provided as inputs using an appropriately-selected standard life table. The method combines two different 1-parameter model life tables, one for children that exactly fits the observed measure of under-five mortality (${}_5q_0$) and one for adults that exactly fits the observed index of conditional survivorship in adulthood (e.g., ${}_{45}q_{15}$ or ${}_{35}q_{15}$).

The model life tables for adults and children are spliced together at a boundary age of 15 years. This age is close to the age at which mortality reaches its minimum in most life tables. Thus, each of the two models represents one arm of the age-specific mortality schedule. Moreover, exact age 15 is conventionally taken to represent the dividing point between child and adult mortality.

***Step 1.1: Derive an estimate of $\alpha$ for the child segment of the life table***

From the estimate of ${}_5q_0$, the value of $\alpha$ for the child segment of the life table (that is, up to age 15) is derived using the relationship

$$\alpha^{child} = Y(5) - Y^s(5),$$

where $Y(5)$ is the logit of the observed ${}_5q_0$,

$$Y(5) = \frac{1}{2}\ln\left(\frac{{}_5q_0}{1-{}_5q_0}\right),$$

and $Y^s(5)$ is the equivalent logit from the selected standard life table, denoted with a superscript *s*.

***Step 1.2: Derive an estimate of $\alpha$ for the adult segment of the life table***

The estimate of $\alpha$ for adults is derived using the relationship

$$\begin{aligned}\alpha^{adult} &= \text{logit}({}_np_{15}) - \text{logit}({}_np^s_{15}) \\ &= \frac{1}{2}\ln\left(\frac{{}_nq_{15}}{1-{}_nq_{15}}\right) - \frac{1}{2}\ln\left(\frac{1-{}_np^s_{15}}{{}_np^s_{15}}\right),\end{aligned}$$

where the first term is calculated from the observed measure of adult mortality conditional on survival to age 15, ${}_nq_{15}$, and the second term is calculated using the identity

$${}_np^s_{15} = \frac{l^s_{15+n}}{l^s_{15}} = \frac{\left(1+\exp(2Y^s(15+n))\right)^{-1}}{\left(1+\exp(2Y^s(15))\right)^{-1}} = \frac{1+\exp\left(2Y^s(15)\right)}{1+\exp\left(2Y^s(15+n)\right)}.$$

Note that, because $\alpha^{adult}$ is calculated for a life table with its radix at age 15, it cannot be directly compared with $\alpha^{child}$.

***Step 1.3: Splice the two segments of the life table together***

In order to splice together a life table using age 15 as the knot, it is necessary to derive the logits of the standard life table's conditional probabilities of survival from age 15, ${}_ap^s_{15}$, for $a$ = 5, 10, 15… Using the formula for ${}_ap^s_{15}$ derived in the previous section to obtain ${}_ap^s_{15}$, these are calculated as

$$\begin{aligned}\text{logit}({}_ap^s_{15}) &= -\frac{1}{2}\ln\left(\frac{{}_ap^s_{15}}{1-{}_ap^s_{15}}\right) \\ &= -\frac{1}{2}\ln\left(\frac{\left(1+\exp\left(2Y^s(15)\right)\right)\Big/\left(1+\exp\left(2Y^s(15+a)\right)\right)}{1-\left(1+\exp\left(2Y^s(15)\right)\right)\Big/\left(1+\exp\left(2Y^s(15+a)\right)\right)}\right).\end{aligned}$$

The final life table, with a radix of 1, is then derived as follows:

For ages less than or equal to 15

$$l(x) = \frac{1}{1+\exp\left(2\left(\alpha^{child} + Y^s(x)\right)\right)}.$$

At age 15 or more (that is, ages 15+$a$ where $a$ = 5, 10, 15, …), the estimated life table survivors at age 15 ($l(15)$, calculated using the $\alpha$ values for children) are multiplied by the fitted conditional probabilities of survivorship from age 15 to give unconditional estimates of survivorship as follows:

$$l(15+a) = l(15)\cdot {}_ap_{15} = \frac{l(15)}{1+\exp\left(2\left(\alpha^{adult} + \text{logit}({}_ap^s_{15})\right)\right)}.$$

Once one has calculated $l(x)$, the other life table functions (e.g. ${}_nm_x$, ${}_nL_x$ and $e_x$) can be calculated using appropriate separation factors, ${}_na_x$.

## METHOD 2: THE BRASS LOGIT METHOD

Chapter 31 introduced the Brass relational logit model life table system in which, with two parameters, $\alpha$ and $\beta$,

$$l(x) = \frac{1}{1+\exp\left(2\left(\alpha + \beta Y^s(x)\right)\right)}.$$

The method for fitting Brass logit models estimates the parameters $\alpha$ and $\beta$ in a relational model life table relative to an appropriately-selected standard that reproduce exactly the observed values of under-five mortality (${}_5q_0$) and the observed index of adult mortality (e.g., ${}_{45}q_{15}$ or ${}_{35}q_{15}$). The two observed measures are assumed to apply to the same point in time.

***Step 2.1: Estimate $\alpha$ and $\beta$***

Because this is a 2-parameter model, it is possible to express either parameter as a function of the other parameter and

one of the two observed measures of mortality. In order to simplify the process of fitting the model, it is useful to express $\alpha$ as a function of $\beta$ and ${}_5q_0$. In the relational logit system of models,

$$Y(5) = 0.5 \ln\left(\frac{{}_5q_0}{1-{}_5q_0}\right)$$

$$Y(5) = \alpha + \beta Y^s(5)$$

and, therefore,

$$\alpha = Y(5) - \beta Y^s(5). \quad (1)$$

Assuming the index of adult mortality is a conditional probability that a 15-year old dies before exact age 15+$n$, ${}_nq_{15}$, then

$${}_nq_{15} = 1 - {}_np_{15} = 1 - \frac{l_{15+n}}{l_{15}} = 1 - \frac{1+\exp\left(2\left(\alpha+\beta Y^s(15)\right)\right)}{1+\exp\left(2\left(\alpha+\beta Y^s(15+n)\right)\right)}.$$

Substituting Equation 1 for $\alpha$, one obtains

$${}_nq_{15} = 1 - \frac{1+\exp\left(2\left(Y(5)+\beta\left(Y^s(15)-Y^s(5)\right)\right)\right)}{1+\exp\left(2\left(Y(5)+\beta\left(Y^s(15+n)-Y^s(5)\right)\right)\right)}.$$

As $Y(5)$ is known, with a tabulated standard it is possible to solve this equation by trial-and-error or by using *Solver* in Excel for the value of $\beta$ that reproduces the observed value of ${}_nq_{15}$. Ideally, the fitted $\beta$ should remain fairly close to its central value of 1. If $\beta$ differs greatly from 1 (e.g. outside the range $0.8<\beta<1.25$), it is advisable to repeat the analysis with an alternative standard life table that has an age pattern of mortality which more closely resembles that of the population in question.

***Step 2.2: Derive a complete life table***

The solution for $\beta$ defines the value of $\alpha$ (through substitution into Equation 1). Using these two parameters and the selected standard, it is possible to generate an entire life table with a radix of 1 using the usual logit relationship

$$l(x) = \frac{1}{1+\exp\left(2\left(\alpha+\beta Y^s(x)\right)\right)}.$$

Once one has calculated $l(x)$, the other life table functions (e.g. ${}_nm_x$, ${}_nL_x$ and $e_x$) can be calculated using appropriate separation factors, ${}_na_x$.

## METHOD 3: THE MODIFIED LOGIT METHOD

The modified logit method, described by Murray, Ferguson, Lopez *et al.* (2003), is a relatively simple extension of the Brass logit system. They proposed a means of deriving a life table using a global standard life table and additional age-specific coefficients $\gamma(x)$ and $\theta(x)$ in a modified logit system

$$Y(x) = \alpha + \beta Y^s(x) + \gamma(x)\left(1 - \frac{Y(5)}{Y^s(5)}\right) + \theta(x)\left(1 - \frac{Y(60)}{Y^s(60)}\right). \quad (2)$$

As before, $Y(x)$ denotes the logit transform of $l(x)$, so the first two elements are those of the Brass system of logit model life tables. The first of the two additional sets of coefficients, $\gamma(x)$ is parameterized by the level of under-five survivorship relative to the standard, while the $\theta(x)$ coefficients are parameterized by survivorship to age 60 relative to the standard. The values of $\gamma(x)$, $\theta(x)$ and the global standard life tables, for males and females, are presented in Table 32.1.

Users of this system of models should be alert to the fact that the values of $\gamma(x)$ and $\theta(x)$ published in the 2003 paper are reversed with respect to sign. Therefore, the coefficients tabulated in Table 3 of that paper should be multiplied by –1 before using them. This correction has been made to the coefficients in Table 32.1.

Despite superficially appearing to be a 4-parameter model, this modification of the logit system of models actually remains a 2-parameter one. Because $\gamma(5)$, $\theta(5)$, $\gamma(60)$ and $\theta(60)$ are all zero, $\alpha$ and $\beta$ fully define $l(5)$ and $l(60)$ in the fitted model and, thereby, the two sets of age-specific deviations from the standard pattern of mortality, $\gamma(x)$ and $\theta(x)$. Thus, the model can be fitted iteratively to any pair of estimates of child mortality and conditional survivorship in adulthood. In particular, the value of $Y(60)$ that defines $Y(x)$ for $x \neq 5,60$ is that in the final fitted life table. It is not necessary, therefore, to know $l(60)$ in advance of fitting the model and a modified logit model can be fitted using any suitable index of adult mortality.

***Step 3.1: Estimate $\alpha$ and $\beta$***

The estimate of $\alpha$ produced by this method can be expressed in terms of $\beta$ and the observed estimate of child mortality, ${}_5q_0$:

$$\alpha = Y(5) - \beta Y^s(5)$$

where $Y(5)$ is the logit of the observed ${}_5q_0$ and $Y^s(5)$ is the logit of $l(5)$ from the standard life table presented in Table 32.1. Moreover, in the fitted life table

$$Y(60) = \alpha + \beta Y^s(60) = Y(5) + \beta\left(Y^s(60) - Y^s(5)\right).$$

Thus, the model can be rewritten in terms of $\beta$, $Y(5)$ and values taken from the standard

$$Y(x) = Y(5) + \beta\left(Y^s(x) - Y^s(5)\right) + \gamma(x)\left(1 - \frac{Y(5)}{Y^s(5)}\right) + \theta(x)\left(1 - \frac{Y(5) + \beta\left(Y^s(60) - Y^s(5)\right)}{Y^s(60)}\right).$$

Using this equation, $\beta$ can be estimated iteratively, with the aim of identifying the value of it that produces a fitted life table that has an index of adult mortality (e.g., ${}_{45}q_{15}$ or ${}_{35}q_{15}$) equal to that originally observed.

***Step 3.2: Derive a complete life table***

The fitted value of $\beta$ and $Y(5)$ can be used to calculate $\alpha$ by means of Equation 1. As $\gamma(60) = \theta(60) = 0$, one can then calculate logit survivorship at age 60

$$Y(60) = \alpha + \beta Y^s(60).$$

With these four items of information ($\alpha, \beta$, $Y(5)$ and $Y(60)$), it is possible to calculate the entire series of $Y(x)$ values using Equation 2 and the coefficients and standards in Table 32.1.

Having obtained $Y(x)$, the associated life table is derived using the usual logit relationship,

$$l(x) = \frac{1}{1 + \exp\left(2Y(x)\right)}.$$

Once one has calculated $l(x)$, the other life table functions (e.g. ${}_nm_x$, ${}_nL_x$ and $e_x$) can be calculated using appropriate separation factors, ${}_na_x$.

*TABLE 32.1 Coefficients, $\gamma(x)$ and $\theta(x)$, and the global standard life table, of the modified logit system of model life tables*

| | Males | | | Females | | |
|---|---|---|---|---|---|---|
| Age ($x$) | $\gamma(x)$ | $\theta(x)$ | $l^s(x)$ | $\gamma(x)$ | $\theta(x)$ | $l^s(x)$ |
| 0 | 0.0000 | 0.0000 | 100,000 | 0.0000 | –0.0000 | 100,000 |
| 1 | –0.1607 | 0.0097 | 96,870 | –0.0855 | –0.0734 | 97,455 |
| 5 | 0 | 0 | 96,010 | 0 | 0 | 96,651 |
| 10 | 0.0325 | –0.0025 | 95,666 | 0.0026 | 0.0229 | 96,370 |
| 15 | 0.0297 | –0.0047 | 95,385 | –0.0291 | 0.0485 | 96,153 |
| 20 | –0.0427 | –0.0018 | 94,782 | –0.1199 | 0.1090 | 95,795 |
| 25 | –0.1262 | 0.0210 | 93,915 | –0.1931 | 0.1702 | 95,340 |
| 30 | –0.1877 | 0.0518 | 93,007 | –0.2352 | 0.2117 | 94,824 |
| 35 | –0.2430 | 0.0883 | 91,949 | –0.2686 | 0.2408 | 94,197 |
| 40 | –0.2899 | 0.1248 | 90,575 | –0.3003 | 0.2601 | 93,370 |
| 45 | –0.3148 | 0.1482 | 88,645 | –0.3203 | 0.2594 | 92,220 |
| 50 | –0.2888 | 0.1402 | 85,834 | –0.2935 | 0.2183 | 90,569 |
| 55 | –0.1915 | 0.0910 | 81,713 | –0.1967 | 0.1338 | 88,159 |
| 60 | 0 | 0 | 75,792 | 0 | 0 | 84,679 |
| 65 | 0.2304 | –0.1170 | 67,493 | 0.2794 | –0.1859 | 79,481 |
| 70 | 0.5523 | –0.2579 | 56,546 | 0.7066 | –0.4377 | 71,763 |
| 75 | 0.9669 | –0.4150 | 42,989 | 1.2835 | –0.7534 | 60,358 |
| 80 | 1.5013 | –0.5936 | 28,117 | 2.0296 | –1.1360 | 44,958 |
| 85 | 2.2126 | –0.8051 | 14,364 | 2.9576 | –1.5774 | 27,123 |

Source: Murray, Ferguson, Lopez *et al.* (2003, Table 3, with the signs of $\gamma$ and $\theta$ reversed (see text for an explanation))

## METHOD 4: THE LOG-QUADRATIC METHOD

An alternative approach to deriving life tables from limited data was recently published by Wilmoth, Zureick, Canudas-Romo *et al.* (2012). It models the age-specific death rates (${}_nm_x$) in a population as a function of age-specific scalar constants $a(x)$, $b(x)$, $c(x)$ and $v(x)$ and parameters $h$ and $k$ in the following relationship

$$\ln\left({}_nm_x\right) = a(x) + b(x)h + c(x)h^2 + v(x)k\,.$$

Values of the four scalar constants were derived from the mortality data contained in the Human Mortality Database, leaving the system with two parameters ($h$ and $k$) used to fit the model. The values of $a(x)$, $b(x)$, $c(x)$ and $v(x)$, for males and females, are presented in Table 32.2.

The parameter $h$ is defined to be the log of the observed ${}_5q_0$, while $k$ is a parameter that (in combination with $v(x)$) describes the deviation of the observed age pattern of mortality from that of a standard population. In practice, it is set to reproduce the observed index of adult mortality conditional on the observed ${}_5q_0$.

### *Step 4.1: Estimate h and k*

The estimate of $h$ in this method is derived from the observed estimate of child mortality, ${}_5q_0$

$$h = \ln({}_5q_0).$$

The value of $k$ is derived by iteratively changing its value to raise or lower the age-specific death rates until the

*TABLE 32.2 Coefficients, a(x), b(x), c(x) and v(x) of the log-quadratic system of model life tables*

| | | Males | | | | Females | | | |
|---|---|---|---|---|---|---|---|---|---|
| *x* | *n* | *a*(*x*) | *b*(*x*) | *c*(*x*) | *v*(*x*) | *a*(*x*) | *b*(*x*) | *c*(*x*) | *v*(*x*) |
| 0 | 1 | –0.5101 | 0.8164 | –0.0245 | 0 | –0.6619 | 0.7684 | –0.0277 | 0 |
| 1 | 4 | | | | | | | | |
| 5 | 5 | –3.0435 | 1.5270 | 0.0817 | 0.1720 | –2.5608 | 1.7937 | 0.1082 | 0.2788 |
| 10 | 5 | –3.9554 | 1.2390 | 0.0638 | 0.1683 | –3.2435 | 1.6653 | 0.1088 | 0.3423 |
| 15 | 5 | –3.9374 | 1.0425 | 0.0750 | 0.2161 | –3.1099 | 1.5797 | 0.1147 | 0.4007 |
| 20 | 5 | –3.4165 | 1.1651 | 0.0945 | 0.3022 | –2.9789 | 1.5053 | 0.1011 | 0.4133 |
| 25 | 5 | –3.4237 | 1.1444 | 0.0905 | 0.3624 | –3.0185 | 1.3729 | 0.0815 | 0.3884 |
| 30 | 5 | –3.4438 | 1.0682 | 0.0814 | 0.3848 | –3.0201 | 1.2879 | 0.0778 | 0.3391 |
| 35 | 5 | –3.4198 | 0.9620 | 0.0714 | 0.3779 | –3.1487 | 1.1071 | 0.0637 | 0.2829 |
| 40 | 5 | –3.3829 | 0.8337 | 0.0609 | 0.3530 | –3.2690 | 0.9339 | 0.0533 | 0.2246 |
| 45 | 5 | –3.4456 | 0.6039 | 0.0362 | 0.3060 | –3.5202 | 0.6642 | 0.0289 | 0.1774 |
| 50 | 5 | –3.4217 | 0.4001 | 0.0138 | 0.2564 | –3.4076 | 0.5556 | 0.0208 | 0.1429 |
| 55 | 5 | –3.4144 | 0.1760 | –0.0128 | 0.2017 | –3.2587 | 0.4461 | 0.0101 | 0.1190 |
| 60 | 5 | –3.1402 | 0.0921 | –0.0216 | 0.1616 | –2.8907 | 0.3988 | 0.0042 | 0.0807 |
| 65 | 5 | –2.8565 | 0.0217 | –0.0283 | 0.1216 | –2.6608 | 0.2591 | –0.0135 | 0.0571 |
| 70 | 5 | –2.4114 | 0.0388 | –0.0235 | 0.0864 | –2.2949 | 0.1759 | –0.0229 | 0.0295 |
| 75 | 5 | –2.0411 | 0.0093 | –0.0252 | 0.0537 | –2.0414 | 0.0481 | –0.0354 | 0.0114 |
| 80 | 5 | –1.6456 | 0.0085 | –0.0221 | 0.0316 | –1.7308 | –0.0064 | –0.0347 | 0.0033 |
| 85 | 5 | –1.3203 | –0.0183 | –0.0219 | 0.0061 | –1.4473 | –0.0531 | –0.0327 | 0.0040 |
| 90 | 5 | –1.0368 | –0.0314 | –0.0184 | 0 | –1.1582 | –0.0617 | –0.0259 | 0 |
| 95 | 5 | –0.7310 | –0.0170 | –0.0133 | 0 | –0.8655 | –0.0598 | –0.0198 | 0 |
| 100 | 5 | –0.5024 | –0.0081 | –0.0086 | 0 | –0.6294 | –0.0513 | –0.0134 | 0 |
| 105 | 5 | –0.3275 | –0.0001 | –0.0048 | 0 | –0.4282 | –0.0341 | –0.0075 | 0 |
| 110 | | –0.2212 | –0.0028 | –0.0027 | 0 | –0.2966 | –0.0229 | –0.0041 | 0 |

Source: Wilmoth, Zureick, Canudas-Romo *et al.* (2012, Table 3)

conditional survivorship ratio in the fitted model life table (e.g., ${}_{45}q_{15}$ or ${}_{35}q_{15}$) matches the observed estimate of the same measure.

***Step 4.2: Derive a complete life table***

Once $h$ and $k$ have been estimated, it is straightforward to calculate a complete series of estimates of ${}_5m_x$ using the coefficients in Table 32.2. Note that, in order to ensure that the fitted estimate of ${}_5q_0$ matches that observed, mortality at ages 1 to 4 is calculated as a residual from the original estimate of ${}_5q_0$ and the fitted infant mortality rate, ${}_1q_0$,

$$ {}_4q_1 = 1 - \frac{1 - {}_5q_0}{1 - {}_1q_0}. $$

As the log-quadratic models define ${}_nm_x$, the age-specific death rates between ages $x$ and $x+5$, separation factors (${}_na_x$) measuring the average number of years lived in an age interval by those that die in it are required to convert the death rates into probabilities of dying and calculate measures such as ${}_{45}q_{15}$ or ${}_{35}q_{15}$. A simple assumption that usually gives reasonable results for the age range 5–59 years is that those who die in a five-year interval die on average, 2.7 years through that interval. (This assumption is made in the Microsoft Excel workbook associated with this chapter, see website, and progressively smaller separation factors are used at older ages).

More care is needed when determining what separation factor to use to estimate ${}_1q_0$ from ${}_1m_0$ as the assumption made can make a material difference to the answer obtained. In the absence of empirical evidence as to what constitutes an appropriate value for ${}_1a_0$ in the population concerned, one can estimate a plausible value using the equations presented by Preston, Heuveline and Guillot (2001: 48) based on the Princeton West series of model life tables. (These equations are used to estimate ${}_1a_0$ in the associated Microsoft Excel workbook, see website).

## WORKED EXAMPLE

As mentioned in the introduction to this chapter, it would be laborious to present detailed step-by-step examples of the methods as an iterative series of calculations is required to derive the fitted life tables. The associated workbook provides an example of the application of each method to estimates derived from an analysis of mortality in Kenya in the mid-1980s. (Note that these estimates predate the onset of demographically significant mortality from AIDS in Kenya and therefore one can derive life tables for this population using models that fail to reflect the impact that a severe HIV epidemic has on the age pattern of mortality).

The pairs of estimates of child and adult mortality to which the four models were fitted are shown in Table 32.3.

*TABLE 32.3 Estimates of under-five and adult mortality, by sex, in Kenya in the mid-1980s*

| | Males | Females |
|---|---|---|
| ${}_5q_0$ | 0.1180 | 0.1080 |
| ${}_{45}q_{15}$ | 0.2352 | 0.1581 |

Figure 32.1 plots ${}_nm_x$ (on a log scale) in the fitted model life tables based, in the case of the first two methods, on Princeton West standard life tables. Other measures that shed light on the differences between the models are presented in Table 32.4.

It can be seen from Figure 32.1 that the death rates in the spliced and log-quadratic models for men are almost identical. The modified logit model also has very similar death rates to these two models in adulthood, but much lower mortality at ages 5–24. The Brass logit model, on the other hand, agrees well with the spliced and log-quadratic models at young ages but has much lower mortality above about age 60 because of the low value of $\beta$. The characteristics of the four fitted models for women are similar to those of the models for men except that the log-quadratic model has much lower mortality at ages 10–34 than any of the other models.

Table 32.4 illustrates the implications of these differences in the age-specific death rates. The fitted Brass logit models have markedly lower mortality in old age than the other models and appreciably higher life expectancies at birth. On the other hand, the two fitted log-quadratic models have the highest mortality in old age and the lowest life expectancies. The models all yield rather similar estimates of overall survivorship during childhood as a whole (i.e. to age 15). However, the estimates of infant mortality vary materially between the models. Those in the Brass logit models are highest, reflecting the same low values of $\beta$ that led to low mortality in old age. The substantive implication of these differences between the fitted models is that one cannot hope to estimate either the infant mortality rate or mortality in old age accurately except by measuring them accurately using direct methods.

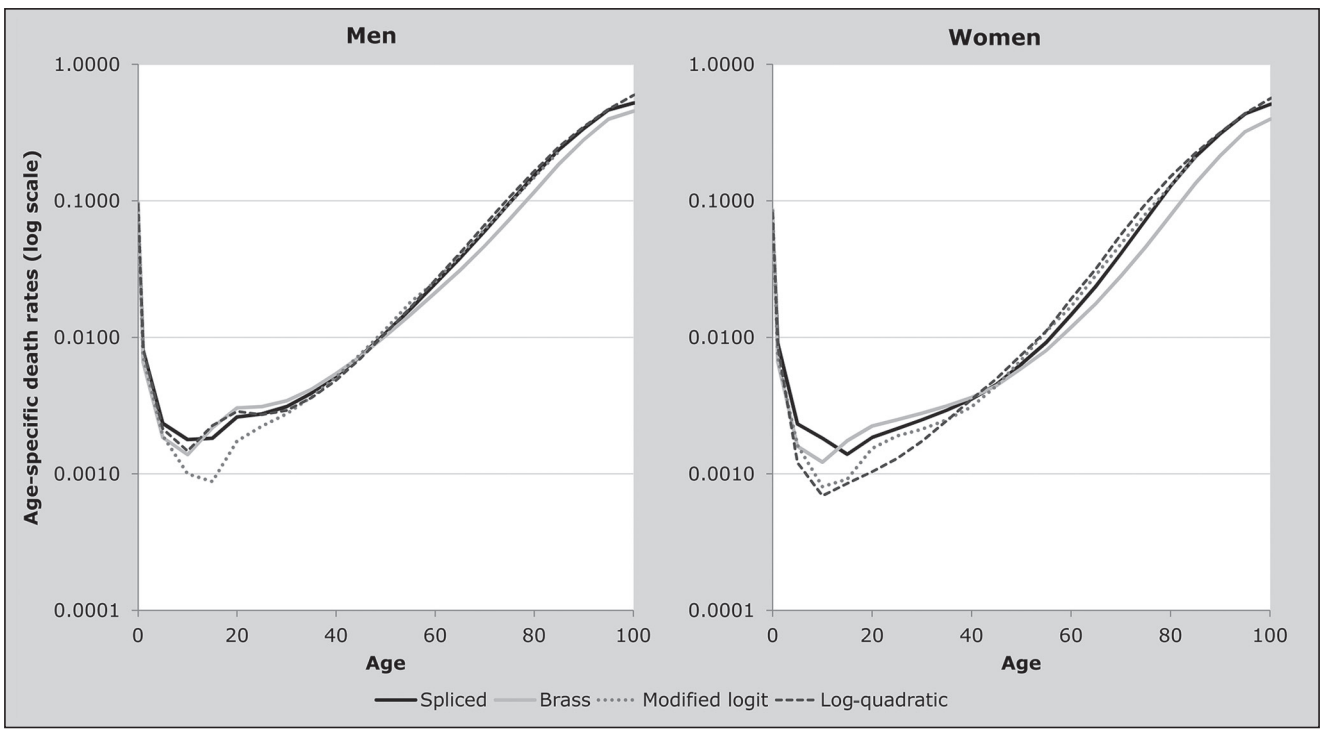
It should be borne in mind that if the first two models were based on a different standard this would change the characteristics of the fitted life tables relative to the modified logit and log-quadratic models. For example, a Princeton South or UN South Asian standard would produce fitted models using the splicing method and Brass logit system of models that match those from the latter two approaches more closely.

While rather few grounds exist to choose one of the methods over the others, the modified logit and log-quadratic models are theoretically superior to the older approaches to fitting a model life table. Moreover, as outlined in the section on Caveats and Warnings, some empirical evidence exists that suggests that they perform better on average than the splicing method and fitting Brass logit models. Third, in contexts in which it is difficult to determine what standard

*TABLE 32.4 Indices of mortality in four model life tables fitted to the mortality estimates in Table 32.3 (West standard life tables)*

| | Males | | | | Females | | | |
|---|---|---|---|---|---|---|---|---|
| Measure | Splicing | Brass logit | Modified logit | Log-quadratic | Splicing | Brass logit | Modified logit | Log-quadratic |
| $e_0$ | 59.7 | 60.9 | 60.3 | 59.3 | 64.3 | 66.6 | 64.4 | 63.9 |
| ${}_1q_0$ | 0.0891 | 0.0944 | 0.0923 | 0.0880 | 0.0815 | 0.0842 | 0.0823 | 0.0768 |
| $l(15)$ | 0.8640 | 0.8679 | 0.8693 | 0.8662 | 0.8737 | 0.8796 | 0.8812 | 0.8836 |
| ${}_{20}q_{60}$ | 0.6706 | 0.5793 | 0.6771 | 0.7054 | 0.5346 | 0.4050 | 0.5840 | 0.6416 |

*FIGURE 32.1 Estimated values of ${}_nm_x$ (on a log scale) in four model life tables fitted to the mortality estimates in Table 32.3 (West standard life tables)*



life table to adopt, this issue can be sidestepped by adopting one of these approaches. Thus, in general, the modified logit and log-quadratic models are to be preferred to the longer-established methods unless evidence exists supporting use of a method based on standards from a particular family of Princeton or United Nations model life tables.

## REFERENCES


Murray CJ, BD Ferguson, AD Lopez, M Guillot, J Salomon and O Ahmad. 2003. "Modified logit life table system: principles, empirical validation, and application", *Population Studies* **57**(2):165–182. doi: 10.1080/0032472032000097083

Preston SH, P Heuveline and M Guillot. 2001. *Demography: Measuring and Modelling Population Processes.* Oxford: Blackwell.

Wilmoth JR, S Zureick, V Canudas-Romo, M Inoue and C Sawyer. 2012. "A flexible two-dimensional mortality model for use in indirect estimation", *Population Studies* **66**(1):1–28. doi: 10.1080/00324728.2011.611411

# Chapter 33 Combining indirect estimates of child and adult mortality to produce a life table

*Ian M Timæus and Tom A Moultrie*

## DESCRIPTION OF THE METHOD

The indirect methods described in this manual for deriving estimates of child and adult mortality produce series of estimates of child and adult mortality, which – using the time location approach pioneered by Feeney (1980, 1991) for children and Brass and Bamgboye (1981) and Brass (1985) for adults – apply to a variety of dates. In many demographic applications, however, it is useful if one can derive an abridged life table that reflects mortality over the entire age range at a specific date within the period covered by such series of indirect estimates of mortality. These applications include the production of population projections or the evaluation of changes in life expectancies at birth or mortality over time.

A general summary of the nature and range of estimates produced by the most important indirect methods is presented in Table 33.1.

An important feature of the estimates that these methods produce for adults is that they are all conditional estimates of survivorship, that measure survival from one age (e.g., 25 in the case of the maternal orphanhood method) to another age (e.g., 35). One cannot straightforwardly convert these conditional measures of survivorship in adulthood into unconditional ones. Thus, the methods for fitting logit model life tables explained by the introductory descriptions of the models in a number of textbooks cannot be applied and more complicated fitting methods are required.

In order to combine estimates of child and adult mortality into a single life table applicable at a defined point in time, a method is needed which addresses the following list of issues:

- The adult mortality estimates need to be converted from their initial conditional form into measures of survivorship from birth.
- The child and adult mortality estimates may imply different patterns or levels of mortality, different time trends in mortality, or both.
- Some data points may be defective or suffer from random fluctuations that distort the overall trend, which implies that the implied trend may require smoothing or adjustment.
- The estimates of child and adult mortality typically refer to different dates and may span different periods of time.
- Neither the methods for estimating child mortality nor those for estimating adult mortality produce any information on the mortality of some age groups, implying that the one can only produce a complete life table by using models.

The method described here seeks to find the parameters $\alpha$ and $\beta$ of a relational logit model life table (described in Chapter 31) applicable to a specified point in time that

| Method | Measure and typical age range | Typical time reference |
|---|---|---|
| Child: Indirect | $l(1) \ldots l(20)$ | 1 to 15 years before the survey |
| Adult: Maternal orphanhood | ${}_{10}p_{25} \cdots {}_{40}p_{25}$ | 3 to 15 years before the survey |
| Adult: Paternal orphanhood | ${}_{15}p_{35} \cdots {}_{35}p_{35}$ | 5 to 15 years before the survey |
| Adult: Siblinghood method | ${}_{10}p_{15} \cdots {}_{35}p_{15}$ | 3 to 15 years before the survey |

***TABLE 33.1*** *Indices of mortality and time references of the estimates produced by selected indirect methods for the estimation of child and adult mortality*

offers the best fit to the observed data points used as inputs. Fitting a 2-parameter model is only possible if independent estimates are available of child and adult mortality for the date in question. If such data are available, fitting a 2-parameter model is recommended because no justification usually exists *a priori* for making the assumption that the age pattern of mortality in the population in question corresponds to that in any particular 1-parameter family of model life tables.

Starting with the observed quantities from the child and adult estimation, the method first derives and plots the implied values of $\alpha$ (the level parameter of a relational model life table) against the time location of each estimate, separately for child and adult mortality making the assumption that $\beta$ (the shape parameter) is equal to 1. This 'alpha plot' is used to identify which data points describe a coherent and consistent trend in the value of $\alpha$ over time. The selected points are then used to iteratively calculate final estimates of $\alpha$ and $\beta$ at the date for which the life table is required. A fitted model life table can then be calculated from the standard using these values of $\alpha$ and $\beta$. The method allows both the $\alpha$ and $\beta$ parameters of the fitted models to change over time but constrains them to do so linearly (Timæus 1990).

The method can be used to derive abridged life tables from sex-specific estimates of child mortality produced by the indirect method for the analysis of data on women's children ever-born and still alive (Chapter 16), and sex-specific estimates of adult mortality produced by application of either the One Census Orphanhood (Chapter 22) or Indirect Siblinghood (Chapter 23) methods. Estimates of child and adult mortality made by direct methods, or from the application of two-census methods, normally apply to a specific year or period of time. Model life tables can be fitted to pairs of estimates of adult and child mortality that refer to the same calendar time using the methods described in Chapter 32.

## DATA REQUIREMENTS AND ASSUMPTIONS

### Tabulations of data required

- A series of sex-specific indirect estimates of child mortality, with their time locations, derived from data on women's children ever born and surviving
- A series of sex-specific estimates of adult mortality, with their time locations, derived using either the indirect method for analysing data on sibling survival or the one census orphanhood method.

In principle, the approach used to fit such data could be extended to estimate life tables for populations for which multiple overlapping sets of indirect estimates exist describing child and adult mortality. However, the workbook that accompanies this chapter has only been designed to handle two series of estimates: one for children and one for adults.

### Assumptions

The method described here bases the fitted model life table on a standard life table. This standard is assumed to have an age pattern of mortality that resembles that of the population being studied. In particular, the relative severity of child and adult mortality in the standard should be similar to that indicated by the indirect estimates to which the model is being fitted. Guidelines for choosing an appropriate standard life table are provided in Chapter 31, which also describes the basic mechanics of the relational logit system of model life tables. The standard need not be taken from the family of model life tables that underlies the coefficients that were used to produce the indirect estimates of child mortality: the family of models that best represents the age pattern of mortality *within* childhood may not be the family that best represents the relative levels of child and adult mortality in the same population.

## CAVEATS AND WARNINGS

The plausibility of the fitted model life table produced by this method of fitting depends on whether the chosen standard life table is appropriate for the population under study. In populations affected by HIV/AIDS, for example, both the balance between child and adult mortality and the detailed age pattern of mortality differ greatly from those that characterize the systems of model life tables in widespread use. Consequently, this method is not recommended for routine application in these circumstances or to other populations for which no standard life table can be identified that describes the balance between child and adult mortality.

## DESCRIPTION OF METHOD

The method is implemented using the following steps.

### *Step 1: Identify the date to which the desired life table should apply*

To avoid the risks associated with out-of-sample extrapolation in the determination of $\alpha$ and $\beta$, the life table should be

fitted to a date within the period covered by the estimates of adult and child estimates that are being analysed. In the presentation of the method that follows, this target date is denoted by $D$.

The exact date for which the life table is required may be determined by the use to which it is going to be put. Ideally a date should be chosen, however, for which both the estimates of adult and child mortality seem reliable. For example, if either the more distant estimates for children appear to biased downward by underreporting of dead children or the more recent estimates for adults appear to be biased downward by the adoption effect, one should avoid producing a life table for the dates covered by the defective estimates. Unfortunately, such considerations sometimes lead the analyst to the conclusion that the data at hand fail to provide a sound basis for the construction of a life table!

If a life table is needed for a more recent, or possibly a more distant, date than the period of time covered by the estimates, a limited amount of extrapolation beyond this range of dates might be considered. The extent of this should be restricted to three years before the earliest time location of any adult or child mortality estimates, on the one hand, and to three years after the earlier of the most recent estimate of child mortality and the most recent estimate of adult mortality, on the other.

***Step 2: Select a standard to be used to derive the fitted life table***

The associated spreadsheet (see website) allows the analyst to choose between nine sex-specific standards: the five UN model life tables (General; South Asian; Far Asian; Latin American; Chilean) and the four Princeton regional model life tables (North; South; East; West). All the standard life tables have a life expectancy at birth of 60 years. The derivation of these logits is described in Chapter 31, and a spreadsheet containing their values can be downloaded from the *Tools for Demographic Estimation* website.

The primary objective in selecting a standard should be to identify one in which the relationship between child and adult mortality is approximately the same as that indicated by the estimates of child and adult mortality. As a practical rule of thumb, if the value of $\beta$ of the final fitted model lies outside the range 0.75–1.25, one should at least consider adopting another standard. In more extreme circumstances, model life tables in which $\beta$ falls outside the range 0.6–1.4 are unlikely to represent empirical mortality schedules adequately. A secondary objective in choosing a standard should be to identify one that shares other known characteristics of the population in question such as the relationship between infant mortality and mortality at ages 1 to 4. The characteristics of the different Princeton and UN families of model life tables are described briefly in Chapter 31.

***Step 3: Plot values of $\alpha$ (assuming $\beta = 1$) derived from the mortality estimates against time***

When $\beta$ (the shape parameter in a relational model life table system) equals 1, the relational model life table system can be expressed as

$$Y(x) = \alpha + Y^s(x)$$

where $Y(x)$ is the logit transform,

$$Y(x) = \text{logit}\big(l(x)\big) = \frac{1}{2}\ln\left(\frac{1-l(x)}{l(x)}\right) = -\frac{1}{2}\ln\left(\frac{1-q(x)}{q(x)}\right).$$

For child mortality, calculating a series of values of $\alpha$ from the estimates is straightforward. The logits of the chosen standard life table for ages 1, 2, 3, 5, 10, 15 and 20 are subtracted from the logits of the derived estimates of child mortality ($q(1)$, $q(2)$, $q(3)$, $q(5)$, $q(10)$, $q(15)$ and $q(20)$):

$$\alpha^{child} = Y(x) - Y^s(x).$$

For adults, the calculation of $\alpha$ is more complicated as the survival probabilities produced by the estimation methods are conditional on survival to a given base age. The formula for $\alpha$ is

$$\alpha^{adults} = \frac{1}{2}\left\{\ln(1-{}_n p_x) - \ln\left[{}_n p_x . \exp\left(2Y^s(x+n)\right) - \exp\left(2Y^s(x)\right)\right]\right\},$$

where $x$ is the base age of the conditional probability of survival (25 for the maternal orphanhood method) and $n$ is the duration over which survivorship is measured, which is contingent on the age group of the respondent. (The derivation of this expression can be found at the end of this chapter).

The estimates of $\alpha$ (separately for children and adults) are then plotted against their respective time locations on the same set of axes.

***Step 4: Eliminate those points in the alpha plot that appear out of line with the general trend***

In order to estimate $\alpha$ and $\beta$ for a specific point in time, the method imposes a linear trend on both parameters. As the first step to achieving this goal, we would like the plots of each of the series of $\alpha$'s (that is, for children and adults separately) against their time locations derived in Step 3 to lie on straight lines.

The $\alpha$'s for individual data points in a series of child or adult mortality estimates derived using the two formulae above may deviate from a straight line for several reasons. First, the underlying pattern of change in mortality may have been strongly non-linear. This is somewhat unlikely given that the series of estimates cover fairly short periods of time and that indirect estimation methods tend to smooth out short term fluctuations in mortality. Even if the diagnostic plot derived above suggests that it is the case, it may still be possible to obtain an adequate fitted model life table by calculating it for a date at which the linear trends in the parameters imposed by the method intersect with the curve indicated by the plotted points. Second, the series may be rather erratic due to sampling errors and reporting errors such as age misstatement. If this is the only limitation of the estimates, one would normally include them all in the analysis and rely on the line fitting procedure to average across these fluctuations.

Third, indirect estimates are vulnerable to biases resulting from respondents failing to answer the key questions accurately or to breaches in the assumptions of the methods concerned. Likely errors in the estimates are discussed in the chapters on the various methods and the reader is referred to those chapters for advice on diagnostic signs that may suggest indicate certain points are biased and should be dropped from the fitting procedure. It is particularly common, however, for the point relating to respondents aged 15–19 in the child mortality method to be biased upward and for the points relating to children aged 5–14 reporting on the survival of their parents in the one census orphanhood method to be biased downward. It will often be necessary to exclude these data points from the model fitting process.

A fourth possible explanation for the failure of the calculated $\alpha$'s to lie on a straight line is that the standard selected for calculating the original estimates may not have been appropriate. If this is the case, it may be necessary to recalculate these estimates using a different standard. Alternately, it may be necessary to try using an alternative standard (as described in Step 2) to derive the fitted life table.

Once the child and adult $\alpha$'s for inclusion in the fitting process have been selected, the rest of the fitting process proceeds mechanically.

***Step 5: Determine the trend in $\beta$ by iteration***

The process of solving for $\beta$ iteratively is not readily done manually, and the associated workbook has been designed to perform the calculations. In order to enable the iteration routine, ensure that Microsoft Excel has been configured appropriately. This is done by selecting "File → Options → Formulas" and then checking the "Enable iterative calculation" checkbox. Setting a maximum of 1000 iterations and a maximum change of 0.00001 is more than sufficient for a solution to be reached.

The process whereby $\beta$ and $\alpha$ are adjusted iteratively to secure a good fit is described in the section on the Mathematical Exposition of the method. The key constraints placed on the fitting process are as follows:

- No matter what the original values of $x$ and $n$ in the estimates of $q(x)$ and ${}_np_b$ for children and adults respectively at the date in question, $\beta$ is calculated consistently from survivorship from age 15 to 60 relative to the standard.
- Both $\alpha$ and $\beta$ are allowed to change over time but it is assumed that they do so linearly.

In combination, these assumptions reduce the distorting impact that errors in the estimates and minor differences in the age pattern of mortality between the population and the standard can have on the fitted model life table (Timæus 1990). In contrast, if one uses the method described in Chapter 32 to fit a 2-parameter logit model life table to a pair of recent indirect estimates of child and adult mortality that refer to about the same date but only measure mortality over a limited range of ages (Brass 1975, 1985), for example $q(2)$ and ${}_{10}p_{25}$, one frequently obtains extreme values of $\beta$ that produce implausible fitted models.

***Step 6: Examine the resulting fitted values of $\alpha$***

The penultimate step is to examine the alpha plot that results from the iterative fitting procedure, which is presented as the second plot of the *alpha plots* sheet of the associated workbook. It is this plot, which presents estimates of $\alpha$ that have been adjusted for the level and trend in $\beta$, that provides

a check on the assumption that $\alpha$ has followed a linear trend. Moreover, if the standard to which the data have been fitted is appropriate, the series of estimates of childhood and adult mortality should lie close to each other in this plot.

### *Step 7: Production of a fitted life table*

Once the best fitting linear time trends in $\alpha$ and $\beta$ have been identified by the iterative fitting process, fitted values of $\alpha$ and $\beta$ for the date for which a fitted life table is required, $D$, are calculated as follows:

$$\alpha^* = Z(\alpha) + D.S(\alpha)$$
$$\beta^* = Z(\beta) + D.S(\beta).$$

The abridged fitted life table is derived from these values of $\alpha^*$ and $\beta^*$ and the standard life table by means of the formula

$$l^*(x) = \frac{1}{1+\exp\left(2\left(\alpha^* + \beta^*.Y^s(x)\right)\right)}.$$

## WORKED EXAMPLE

The worked example presented here uses data on the female population from the Dominican Republic. The indirect estimates for girls were made from the data on children ever-born and surviving obtained by a DHS conducted in 2002. The indirect estimates for adult women were made from the reports on the survival of mothers from the census conducted in the same year. The input data are presented in Table 33.2.

### *Step 1: Identify the date to which the desired life table should apply*

In the case of the data from the Dominican Republic, the associated spreadsheet permits a life table to be derived for dates lying between the earlier of 1987.99–3 and 1990.51–3 and the earlier of 2001.71+3 and 1999.23+3, which is to say dates between 1984.99 and 2002.23.

In this example, we derive a life table for the Dominican Republic for mid-1997, i.e. 1997.5.

### *Step 2: Select a standard to be used to derive the fitted life table*

Given the geographical source of the data, it is reasonable to assume (at least initially) that the mortality of women in the Dominican Republic follows the age pattern described by the UN Latin American female standard. The logits of the chosen standard life table are presented in Table 33.3.

*TABLE 33.2 Input data for combining child and adult mortality estimates, Dominican Republic*

| Child mortality (2002 DHS) | | | Adult mortality (2002 Census) | | |
|---|---|---|---|---|---|
| $x$ | $q(x)$ | Date | $n$ | ${}_np_{25}$ | Date |
| 1 | 0.0338 | 2001.71 | 10 | 0.9858 | 1999.23 |
| 2 | 0.0429 | 2000.24 | 15 | 0.9801 | 1997.07 |
| 3 | 0.0355 | 1998.48 | 20 | 0.9680 | 1995.13 |
| 5 | 0.0467 | 1996.43 | 25 | 0.9479 | 1993.43 |
| 10 | 0.0619 | 1994.16 | 30 | 0.9214 | 1992.02 |
| 15 | 0.0710 | 1991.52 | 35 | 0.8872 | 1991.00 |
| 20 | 0.0799 | 1987.99 | 40 | 0.8373 | 1990.51 |

*TABLE 33.3 Logits of the UN Latin American female life table with a life expectancy of 60 years*

| Age ($x$) | $Y^s(x) = \text{logit}(l(x))$ |
|---|---|
| 0 | |
| 1 | –1.2375 |
| 2 | –1.1006 |
| 3 | –1.0398 |
| 4 | –1.0046 |
| 5 | –0.9815 |
| 10 | –0.9304 |
| 15 | –0.9054 |
| 20 | –0.8735 |
| 25 | –0.8313 |
| 30 | –0.7828 |
| 35 | –0.7285 |
| 40 | –0.6670 |
| 45 | –0.6005 |
| 50 | –0.5248 |
| 55 | –0.4356 |
| 60 | –0.3230 |
| 65 | –0.1795 |

Note that, if you change the family of life tables from "UN" to "Princeton" or vice versa in the associated workbook, you must force the workbook to recalculate the output by changing the "Recalculate" cell (B3 on the *Method* sheet) from "True" to "False" and back to "True". Failure to do this will produce an error.

***Step 3: Plot values of $\alpha$ (assuming $\beta = 1$) derived from the mortality estimates against their time locations***

Using the data from the Dominican Republic in Table 33.2 and a UN Latin American life table for a standard, the value of $\alpha$ for child mortality when $x = 3$ is derived as follows:

$$\alpha^{child} = -\frac{1}{2}\ln\left(\frac{1-q(3)}{q(3)}\right) - \text{logit}\left(l^s(3)\right)$$
$$= -\frac{1}{2}\ln\left(\frac{1-0.0355}{0.0355}\right) + 1.0398$$
$$= -0.6112.$$

This value of $\alpha$ has a time location of 1998.48, as indicated in Table 33.1. The values of $\alpha$ for the other estimates of child mortality, together with their time locations are derived similarly.

Using the data on adult mortality in Table 33.2 and the same standard, the estimate of the adult $\alpha$ when $n$ is 25 is given by

$$\alpha^{adults} = \frac{1}{2}\left\{\ln(1 - {}_{25}p_{25}) - \ln\left[\begin{array}{l} {}_{25}p_{25}.\exp\left(2Y^s(50)\right) \\ -\exp\left(2Y^s(25)\right)\end{array}\right]\right\}$$
$$= \frac{1}{2}\left\{\ln(1-0.9479) - \ln\left[\begin{array}{l} 0.9479\exp\left(2(-0.5248)\right) \\ -\exp\left(2(-0.8313)\right)\end{array}\right]\right\}$$
$$= -0.5021.$$

This value of $\alpha$ has a time location of 1993.43. The values of $\alpha$ for the other estimates of adult mortality, together with their time locations are derived similarly.

A summary of these estimates of $\alpha$ and their time locations are presented in Table 33.4.

When all the estimates of $\alpha$ based on data on both children and adults are plotted against their time locations, the alpha plot shown in Figure 33.1 results.

***Step 4: Eliminate those points in the alpha plot that are out of line with the general trend***

Chapter 16 explains that the most recent indirect estimate of child mortality, which is based on the reports of women aged 15–19 tends to be biased upward as teenage mothers are a select group with high mortality because, among other reasons, they tend to come from socially disadvantaged backgrounds. This data point is nearly always ignored when inferences are made about the trend in child mortality from indirect estimates and this was done in this application.

The most recent estimate of adult mortality based on children aged 5–9 reporting on the survival of their mothers in the one census orphanhood method also underestimates mortality in many applications. In the Dominican Republic, however it indicates much higher mortality than one would expect given the trend indicated by the other estimates for adult women. This might be the result of severe underreporting of the ages of children or might indicate that the models involved in the estimation process are inappropriate for this population. Either way it was decided to ignore this anomalous estimate. Therefore, the most recent estimate in each series was omitted from the fitting of a trend line to the $\alpha$'s by clearing its respective cell in the *alpha plots* sheet of the associated workbook.

The remaining estimates from the orphanhood method are internally consistent and suggest that adult women's mortality fell rapidly in the Dominican Republic during the 1990s. The child mortality estimates also suggest that mortality was falling, but the more recent estimates are

*TABLE 33.4 Estimates of $\alpha$ and the time location of the estimates, females, Dominican Republic, 2002*

| Female children | | | Female adults | | |
|---|---|---|---|---|---|
| Original index | $\alpha$ | Time location | Original index | $\alpha$ | Time location |
| $q(1)$ | –0.4389 | 2001.71 | ${}_{10}p_{25}$ | –0.5176 | 1999.23 |
| $q(2)$ | –0.4519 | 2000.24 | ${}_{15}p_{25}$ | –0.6183 | 1997.07 |
| $q(3)$ | –0.6112 | 1998.48 | ${}_{20}p_{25}$ | –0.5779 | 1995.13 |
| $q(5)$ | –0.5266 | 1996.43 | ${}_{25}p_{25}$ | –0.5021 | 1993.43 |
| $q(10)$ | –0.4288 | 1994.16 | ${}_{30}p_{25}$ | –0.4567 | 1992.02 |
| $q(15)$ | –0.3803 | 1991.52 | ${}_{35}p_{25}$ | –0.4463 | 1991.00 |
| $q(20)$ | –0.3483 | 1987.99 | ${}_{40}p_{25}$ | –0.4436 | 1990.51 |

somewhat inconsistent with each other. The 3rd and 4th points, which are based on the reports of mothers aged 25–34 years, indicate that the rate of decline in child mortality accelerated in the second half of the 1990s. However, the 2nd estimate, which is based on the reports of women aged 20–24, suggests that it decelerated. In the absence of evidence as to the nature of the errors in the data that have led to these inconsistencies, it was decided to leave all three data points in the analysis.

The final selection of points produces the alpha plot in Figure 33.2. This plot emphasizes the consistency of the 2nd to 7th points for adults and shows that a regression line fitted to the 2nd to 7th points for children not only passes through the middle of the more recent estimates, but fits the three more distant points well.

Note that, in Figure 33.2, the values of $\alpha$ derived from the estimates of the mortality of adults lie below those derived from the estimates of child mortality and diverge from them over time. This means that, relative to the UN Latin American standard, adult mortality in the Dominican Republic in the 1990s was low and was falling more rapidly than child mortality. Thus, the $\beta$ parameter of fitted model life tables for this population will lie below 1 and will decrease over time.

***Step 5: Determine the trend in $\beta$ by iteration***

The spreadsheet iteratively solves for fitted values of both $\alpha$ and $\beta$ for the desired time point (1997.5). The estimates of $\alpha^*$ and $\beta^*$ are –0.658 and 0.849 respectively. These estimates implies that the level of mortality in the Dominican Republic is somewhat lighter than in the UN Latin American standard ($\alpha < 0$), and that mortality is somewhat heavier at younger ages and lighter at older ages ($\beta < 1$) than in this standard. The estimate of $\beta^*$ is close enough to 1 not to raise any concerns about the choice of standard made in Step 2.

***Step 6: Examine the resulting fitted values of $\alpha$***

The penultimate step is to examine the alpha plot that results from the iterative fitting procedure, which is presented as the second plot of the "alpha plots" sheet of the associated workbook.

Figure 33.3 shows that there is now a close correspondence

*FIGURE 33.1 Initial plot of child and adult $\alpha$'s against their time location, Dominican Republic*

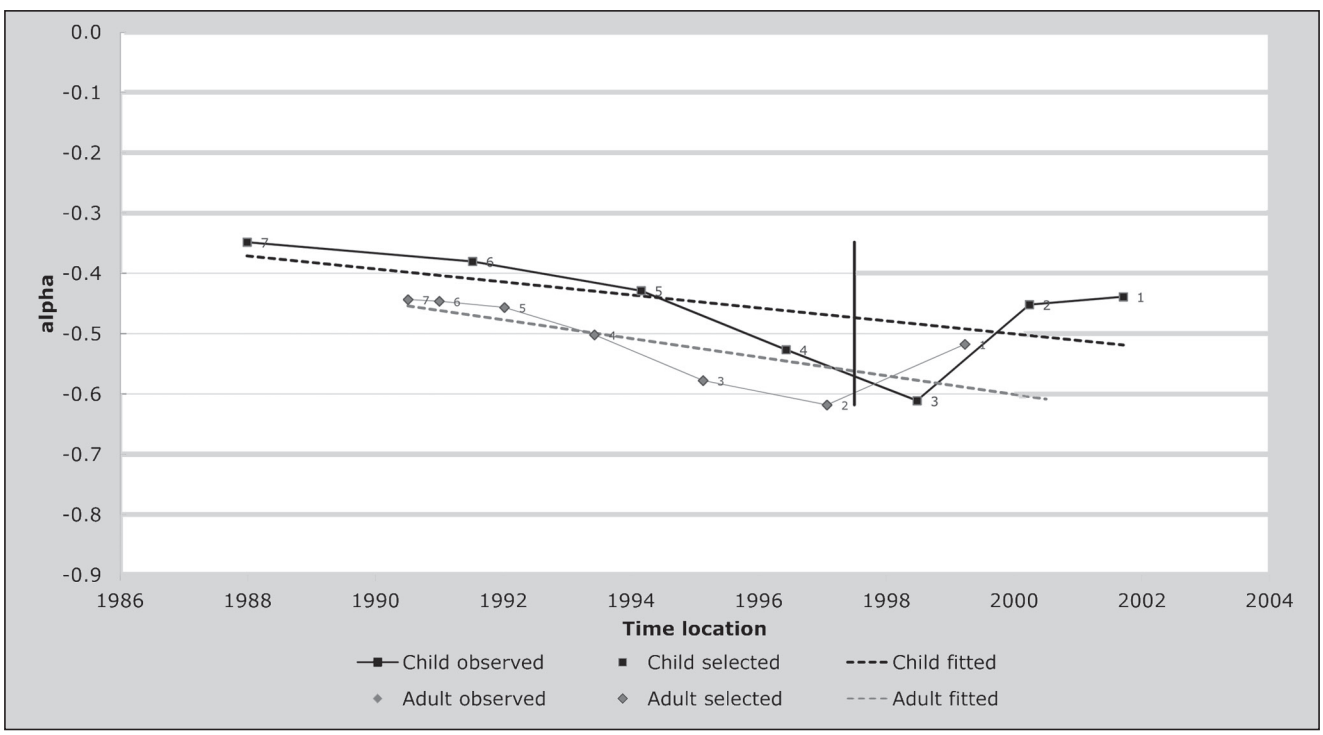


*FIGURE 33.2* *Final plot of child and adult α's against time location, Dominican Republic*

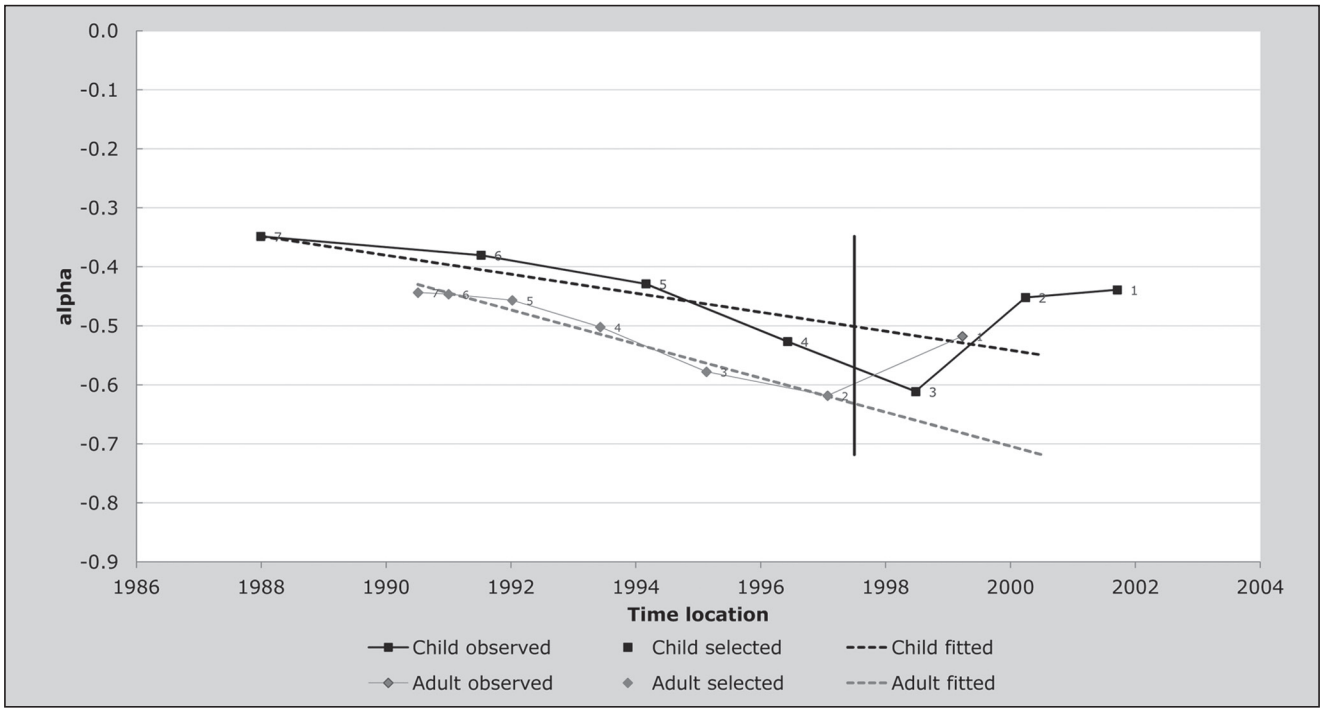
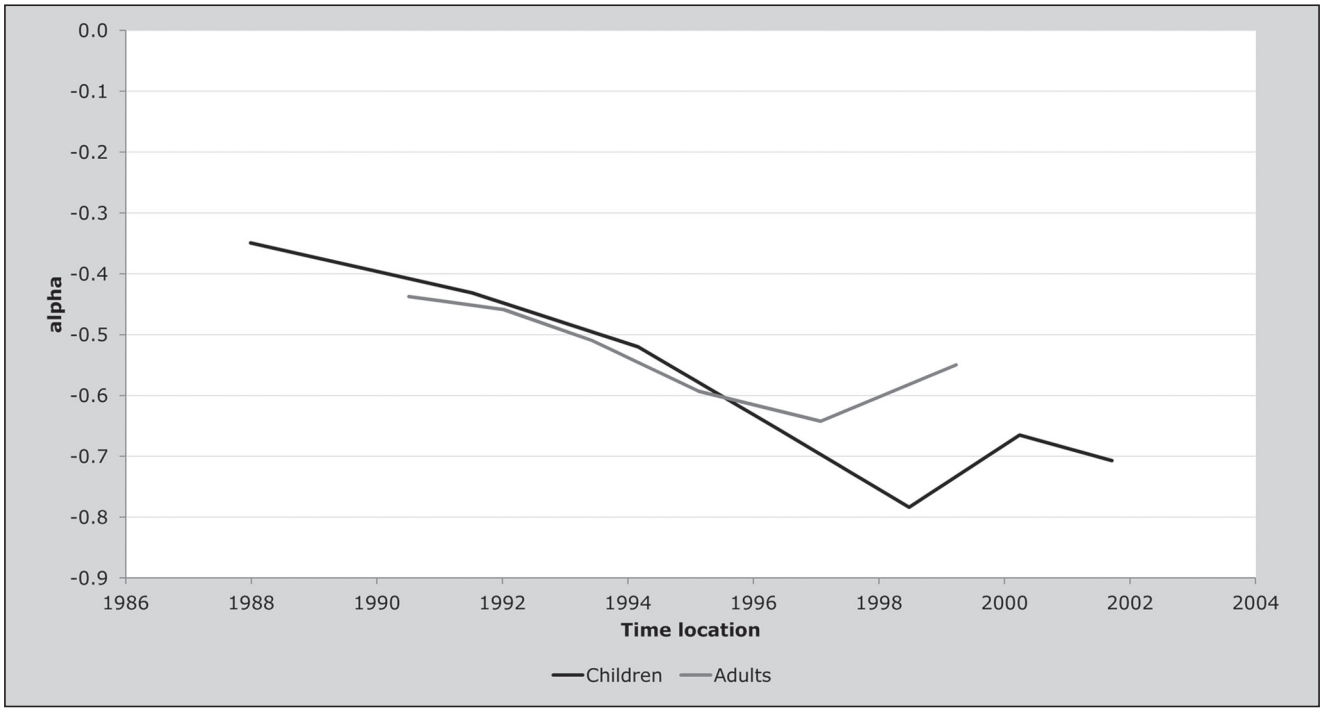



*FIGURE 33.3* *Alpha plot, after fitting β iteratively, Dominican Republic*



between the $\alpha$'s for children and adults for most of the 1990s. Mortality was falling across the age range, though $\beta$ dropped from about 0.95 to 0.85 between the early 1990s and mid-1997. This is what one would expect given that we have already observed that adult mortality in the Dominican Republic was falling rapidly at this time in comparison with the pattern in the family of logit model life tables based on the UN Latin American standard.

The lines for adults and children remain fairly close to each other in 1997, reflecting the fact that the value of $\beta$ at that time (i.e., 0.85) remained fairly close to its central value of 1. The two estimates of $\alpha$ for 1997 would only differ greatly if $\beta$ at this time was very different from 1. If they did differ markedly, it would be advisable to seek out a standard life table that was more appropriate for the population being studied.

If the two series of estimates followed very different trends and failed to cross over each other, or did so and then diverged rapidly, or if one or both series were highly non-linear, this would again suggest that the standard being used was inappropriate or, more probably, that one or both series was severely biased by errors in the data, making it impossible to reconcile them with each other.

***Step 7: Production of a fitted life table***

The abridged fitted life table is derived from the fitted values of $\alpha^* = -0.658$ and $\beta^* = 0.849$ for the selected date, and the standard life table (presented in Table 33.3) by means of the formula

$$l^*(x) = \frac{1}{1+\exp\left(2\left(\alpha^* + \beta^*.Y^s(x)\right)\right)}.$$

The final fitted life table is presented in Table 33.5. Life expectancy at birth is 76.6 years compared with the United Nations' estimate for the same quinquennium of 73.1 years (UN Population Division 2013).

*TABLE 33.5 Fitted life table for females, Dominican Republic, mid-1997*

| Age ($x$) | $l(x)$ |
|---|---|
| 0 | 1.0000 |
| 1 | 0.9683 |
| 2 | 0.9603 |
| 3 | 0.9561 |
| 4 | 0.9536 |
| 5 | 0.9518 |
| 10 | 0.9476 |
| 15 | 0.9455 |
| 20 | 0.9426 |
| 25 | 0.9386 |
| 30 | 0.9337 |
| 35 | 0.9278 |
| 40 | 0.9204 |
| 45 | 0.9118 |
| 50 | 0.9008 |
| 55 | 0.8865 |
| 60 | 0.8657 |
| 65 | 0.8348 |
| 70 | 0.7863 |
| 75 | 0.7120 |
| 80 | 0.6029 |
| 85 | 0.4452 |
| 90 | 0.2587 |
| 95 | 0.1024 |
| 100 | 0.0242 |

## DETAILED DESCRIPTION OF THE METHOD

The associated spreadsheet (see website) implements the method by following the steps outlined above. This section provides a detailed description of how the iterative procedure used to derive the final values of $\alpha$ and $\beta$ is implemented.

The premise underlying the fitting procedure is that the derived life table should fit the observed data well at ages 15 and 60. The former constraint ensures that child and adolescent mortality is well matched; the combination of the two ensures that adult mortality over an extended age range (15 to 60) is close to that implied by the adult mortality estimates used to fit the life table.

**Fitting procedure**

After selecting the data points that will be used (as described in Step 4), the method seeks to find the best fitting linear regression model of the time trend in the estimates of $\alpha$ for children, conditional on the estimated trend in $\beta$, and the best fitting linear regression model of the time trend in the estimates of $\beta$ for adults, conditional on the estimated trend in $\alpha$ for children.

Starting with the assumption that $\beta = 1$, one can calculate an $\alpha^{child}$ corresponding to each estimate of child mortality using the equation provided in Step 3 of the worked example. Since each estimate of $\alpha^{child}$ is associated with its time location ($T$), one can regress the estimates included in

the fitting procedure on time to obtain the slope $S(\alpha)$ and intercept $Z(\alpha)$ of the linear regression model.

The fitted regression model can then be used to predict a fitted $\alpha$ ($\alpha^*$) for the times to which the adult mortality estimates refer

$$\alpha^* = Z(\alpha) + T.S(\alpha).$$

Using these fitted values of $\alpha^{child}$, one can estimate $Y(15)$ at these dates

$$Y(15) = \alpha^* + \beta^* Y^s(15)$$

where, in this first iteration, $\beta^* = 1$.

Still assuming that $\beta = 1$, one can also estimate $\alpha^{adult}$ from the conditional estimates of adult survivorship that have been included in the fitting procedure using the equation given in Step 3 of the worked example and use these values of $\alpha^{adult}$ to calculate corresponding estimates of ${}_{45}q_{15}$. Multiplying the value of $l(15)$ estimated from the data on children by an estimate of ${}_{45}p_{15}$ for the same date estimated from the data on adults gives an unconditional estimate of $l(60)$ and therefore of $Y(60)$:

$$Y(60) = -\frac{1}{2}\ln\left(\frac{l(60)}{1-l(60)}\right) = -\frac{1}{2}\ln\left(\frac{l(15).{}_{45}p_{15}}{1-l(15).{}_{45}p_{15}}\right).$$

The estimate of $l(15)$ is calculated from $Y(15)$ as

$$l(15) = \frac{1}{1+\exp\left(2Y(15)\right)},$$

while that of ${}_{45}p_{15}$ is derived from the $\alpha$'s and $\beta$'s fitted to the adult estimates:

$${}_{45}p_{15} = \frac{l(60)}{l(15)} = \frac{\left(1+\exp\left(2\left(\alpha^{adult} + \beta^*\left(Y^s(15)\right)\right)\right)\right)}{\left(1+\exp\left(2\left(\alpha^{adult} + \beta^*\left(Y^s(60)\right)\right)\right)\right)}$$

where $Y^s(x)$ represents the logit of $l(x)$ in the standard life table (i.e., with $\beta = 1$ and $\alpha = 0$) and, in this first iteration, $\beta^* = 1$.

Having estimated a series of values for $Y(60)$ for the dates to which the adult mortality estimates refer, it is now possible to calculate revised estimates of $\beta$ for these dates

$$\beta = \frac{Y(60) - Y(15)}{Y^s(60) - Y^s(15)}.$$

As these revised $\beta$'s each refer to a specific date, they can then be regressed on time ($T$) to calculate the slope $S(\beta)$ and intercept $Z(\beta)$ of a linear regression line that can then be used to predict a fitted $\beta$ ($\beta^*$) for each data point, whether it is for children or adults, from that point's time location ($T$)

$$\beta^* = Z(\beta) + T.S(\beta).$$

At this point, the first iterative cycle has been completed. One can now calculate revised estimates of $\alpha^{child}$, that allow for the fact that $\beta$ has been allowed to differ from 1, with the formula

$$\alpha^{child} = \text{logit}(q_x) - \beta^*.Y^s(x).$$

The revised estimates of $\alpha^{child}$ are then regressed on time and used in combination the fitted estimates of $\beta$ for the dates to which the adult mortality estimates refer to calculate revised estimates of $Y(15)$, ${}_{45}q_{15}$ and $Y(60)$ and then to recalculate a second round of revised estimates of $\beta$ itself. Thus, we now have in place a mechanism that will iteratively calculate the best fitting regressions of $\alpha$ and $\beta$ on time, each adjusted for the other.

**Derivation of the formula for calculating $\alpha$ for adults**

$${}_np_x = \frac{l(x+n)}{l(x)} = \left[\frac{1+e^{2(\alpha+\beta Y^s(x))}}{1+e^{2(\alpha+\beta Y^s(x+n))}}\right]$$

$${}_np_x + {}_np_x.e^{2(\alpha+\beta Y^s(x+n))} = 1 + e^{2(\alpha+\beta Y^s(x))}$$

$$e^{2\alpha}.({}_np_x.e^{2\beta Y^s(x+n)} - e^{2\beta Y^s(x)}) = 1 - {}_np_x$$

$$e^{2\alpha} = \frac{1-{}_np_x}{{}_np_x.e^{2\beta Y^s(x+n)} - e^{2\beta Y^s(x)}}$$

$$\alpha = \frac{1}{2}\ln\left(\frac{1-{}_np_x}{{}_np_x.e^{2\beta Y^s(x+n)} - e^{2\beta Y^s(x)}}\right)$$

$$\alpha = \frac{1}{2}\left\{\ln\left(1-{}_np_x\right) - \ln\left({}_np_x.e^{2\beta Y^s(x+n)} - e^{2\beta Y^s(x)}\right)\right\}.$$

## REFERENCES


Brass W. 1975. *Methods for Estimating Fertility and Mortality from Limited and Defective Data*. Chapel Hill: International Program of Laboratories for Population Statistics.

Brass W. 1985. *Advances in Methods for Estimating Fertility and Mortality from Limited and Defective Data*. London: London School of Hygiene & Tropical Medicine.

Brass W and EA Bamgboye. 1981. *The Time Location of Reports of Survivorship: Estimates for Maternal and Paternal Orphanhood and the Ever-widowed.* London: London School of Hygiene & Tropical Medicine.

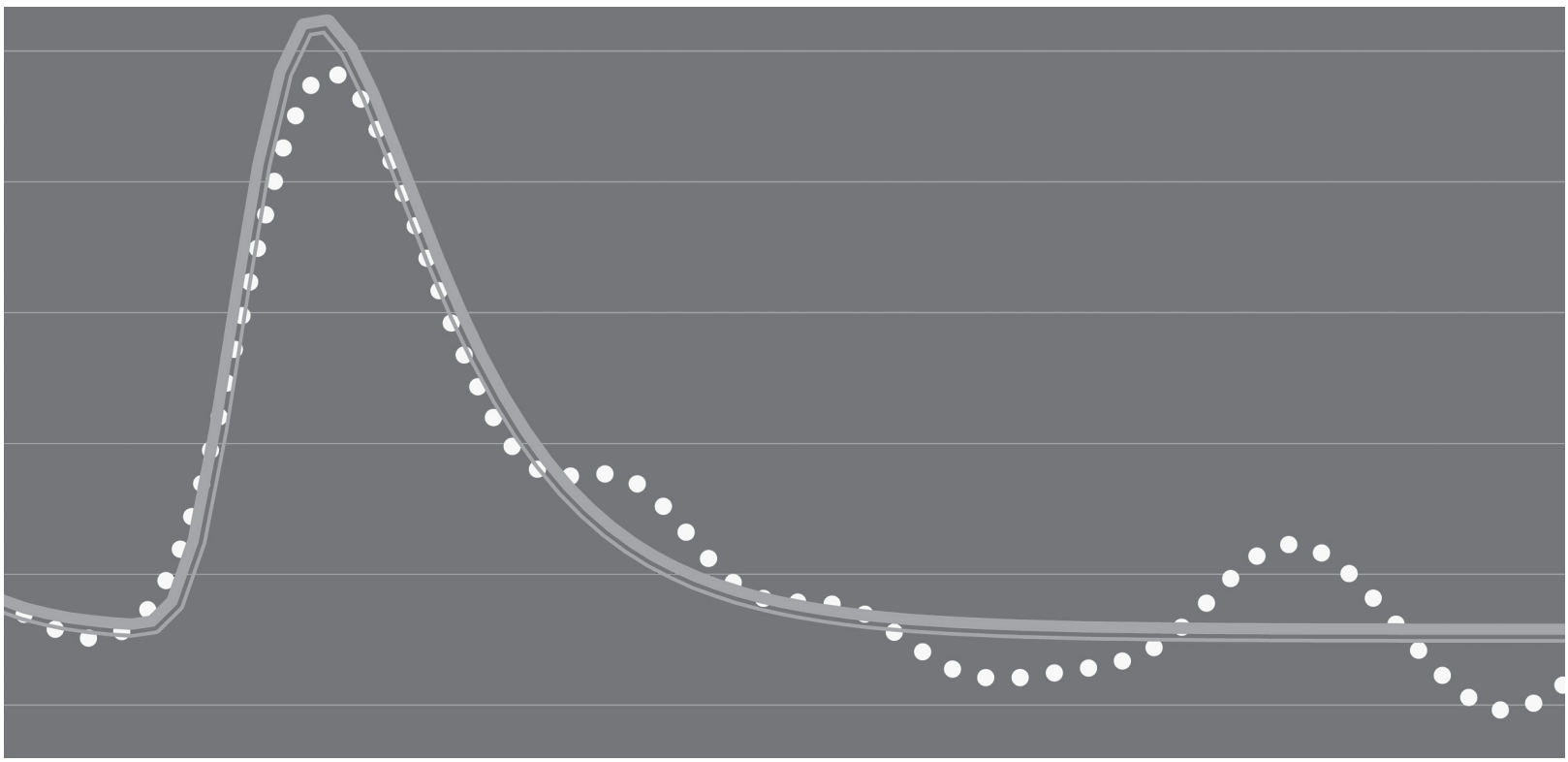
Feeney G. 1980. "Estimating infant mortality trends from child survivorship data", *Population Studies* **34**(1):109–128. doi: 10.1080/00324728.1980.10412839

Feeney G. 1991. "Child survivorship estimation: Methods and data analysis", *Asian and Pacific Population Forum* **5**(2–3):51–55, 76–87. http://hdl.handle.net/10125/3600

Timæus IM. 1990. "Advances in the Measurement of Adult Mortality from Data on Orphanhood." Unpublished PhD thesis, London: University of London.

UN Population Division. 2013. *World Population Prospects: The 2012 Revision*. New York: United Nations, Department of Economic and Social Affairs. http://esa.un.org/wpp/index.htm

# Migration



# Chapter 34 Introduction to migration analysis

*Rob Dorrington and Kenneth Hill*

## OVERVIEW

Migration is the third process (with fertility and mortality) that governs population change. For most national populations, its contribution to population change is small relative to those of births and deaths, but as the civil division of interest becomes smaller, the salience of migration typically becomes larger. Migration differs from fertility and mortality not only in magnitude, but more fundamentally in the nature of the process. Migration involves moving across some geographically-defined boundary, with the intent or result of changing place of normal residence. Thus whereas a birth and a death are largely unambiguous, a migration depends upon geographically-defined spatial units (civil divisions) and on intent or subsequent behaviour. A person can be a migrant to the analyst looking at change in provincial population but not a migrant to another analyst focusing on national population change. The first task, therefore, in any analysis of migration is to establish the geographic focus of the study. A second task is to define what counts as a migration, as opposed to broader mobility. The issue is further confused by the existence of several different types of migration. In addition to "ordinary" change of usual residence, there are circular migration flows, daily or weekly commuter flows, seasonal flows and refugee flows, all with specific characteristics. Given these definitional issues, and the fact that migrations can effectively be reversed in terms of population stocks (unlike births and deaths), it is no surprise that measurement is also complicated.

Apart from this, capturing data on migration is also more problematic. Although developing countries often lack complete systems of birth and death registration, completeness is improving and some methods have been devised to make use of the less than complete data. However, registration data on migrants/migrations in most countries cannot be relied on to produce reliable estimates of immigrants, let alone of internal migrants/migrations. In addition, for various reasons (illegal status, temporary residence of recent migrants, fear of xenophobia, etc.) migrants (especially immigrants) are usually underrepresented in censuses and surveys.

Methods for measuring migration are broadly similar for both internal migration (in- or out-migration) and international migration (immigration or emigration), except in one very important respect. A census or survey can measure international immigration by identifying persons born abroad, but it is much harder to identify emigrants because it is not possible to carry out a census/survey in all recipient countries. Approaches to estimating emigration include: (i) systematic identification of nationals in censuses of other countries (UN Population Division 2011); (ii) including census/survey questions about usual household members living abroad (e.g. in the Swaziland Censuses of 1986 and 1996); (iii) asking about the residence abroad of close relatives, especially a woman's children or a respondent's siblings (Zaba 1985); and (iv) using intercensal residual methods to estimate numbers of missing residents at the time of a second census. The first approach is dependent on receiving countries having, and being willing to share, relevant data and only captures migration of the native-born population; the second approach depends on the, perhaps vague, concept of household membership, and will also fail to cover entire households that have moved away; the third also fails to capture entire missing families, does not provide estimates of recent emigration, and in small experimental surveys has not proven convincing. Only the fourth can be expected to give plausible estimates of recent outflows, provided both censuses count the population reasonably accurately, but gives no potentially useful information about destination.

With these limitations and problems of accurate data

collection, the field of migration analysis has developed largely independently from mainstream demography, leading to it concentrating primarily on developed countries where the quality of data available to measure migration is typically much better than it is in developing countries, and possibly because migration in these countries is often a matter of greater political and public policy concern. A further consequence of these factors is that the field has developed its own terminology and techniques, which are often quite far removed from the demography discussed elsewhere in this manual.

## DEFINITIONS

As noted above, a migration is defined as a move across a geographically-defined (usually administrative) boundary of interest to the analyst with the effect of changing a person's place of usual residence. Assuming that the boundary can be clearly defined, this immediately raises two questions: how does one define usual place of residence, and how does one determine whether it has changed? Unfortunately, no very precise answers can be given to these two questions, giving rise to inevitable uncertainty in measurement. The preferred definition of usual residence is in terms of length of residence: that if one intends to live, or after one has lived, in a place for a period of time (e.g. one year) one becomes a usual resident. Note that usual residence is not the same thing as legal residence. The *Principles and Recommendations for Population and Housing Censuses* (UN Statistics Division 2008: 102, para. 1.463) defines usual residence as follows:

> *"It is recommended that countries apply a threshold of 12 months when considering place of usual residence according to one of the following two criteria:*
>
> *(a) The place at which the person has lived continuously for most of the last 12 months (that is, for at least six months and one day), not including temporary absences for holidays or work assignments, or intends to live for at least six months;*
>
> *(b) The place at which the person has lived continuously for at least the last 12 months, not including temporary absences for holidays or work assignments, or intends to live for at least 12 months."*

However, this definition does not deal with the situation of a person with two homes who regularly spends about six months in each. In general, we have to rely on people to self-define as residents or not, although some tests could be implemented (such as asking where their car is registered, where taxes are paid, where they voted, where the person sleeps at night on a regular basis, etc.). For most purposes, a person can distinguish between whether he or she is a usual resident and visitor, and this simple distinction suffices.

## DATA SOURCES

Migration has been the Cinderella of demography, kept in the background as far as possible, and dedicated migration surveys are few, far between, and specialized (an excellent example is the description of the Mexican Migration Project by Massey, Alarcon, Durand *et al.* (1987)). Dedicated migration surveys typically include full migration histories, which, though raising complex analytical issues, tend not to be focussed on the estimation of numbers of migrants/migrations. In this section we do not cover the analysis of such full histories (there are very few general principles that would apply to a useful number), but rather deal with the sorts of data collected by population censuses and general household surveys and sometimes, developed countries, by some form of registration.

### Birthplace

The most widely collected data relevant to migration is place of birth. In comparison with place of residence at the time of a survey, this information describes lifetime migration. The information provides limited information about timing of migration, and is 'net' migration in the sense that it misses, entirely, migrations that have been reversed (back to the place of birth) and all intermediate migrations. At the time of data collection, decisions have to be taken about the granularity of the data: i.e., for those born abroad, how many countries should be explicitly recorded and for those born in the country, what level of geography should be recorded. For the analyst, of course, these decisions were made at the questionnaire design stage, but some degree of greater aggregation may be required. The analysis of data on birthplace is described below, but it is useful to make two points here. First, if data on birthplace by age and sex are available for two points in time, it is possible to estimate net migration (by age and sex) during the interval. Second, although birthplace reflects lifetime migration, the length of "lifetime" varies by age, and (provided the census data on children is reasonably accurate, which it often isn't in many developing countries) the migration of 0-4 year olds may be used as an indicator for recent migration of their parents (Raymer and Rogers 2007).

**Residence at some specified time in the past**

This information is very often collected in addition to that on birthplace, with the express objective of providing data on recent migration. The time point specified is generally five years earlier, but sometimes a one year period is used. However, it tends to work better if the time point is associated with a memorable event, such as the previous census, on the assumption that the coverage of that previous census was largely complete (so that people remember being counted). The longer time period identifies more migrants, but misses intermediate moves, whereas the shorter time period is more susceptible to reference period error (I moved "about a year ago").

**Place of previous residence**

This information is almost always collected as an alternative to residence at some specified time in the past, and is generally combined with an additional question about duration of current residence (or date of last move). The objective again is to provide data on recent migration.

**Duration of current residence**

The question refers to duration of residence in the civil division (such as a town or province), not in an individual dwelling unit. This question is of limited use on its own and tends to be paired with the one above to provide a time frame for estimates.

**Intercensal population change**

Though not involving a direct question about migration, intercensal population change by age and sex can, provided both censuses are reasonably accurate counts of the population, provide residual estimates of net migration between the two censuses (Hill 1987; Hill and Wong 2005; UN Population Division 1967). Intercensal population change (for cohorts or age groups) by age and sex is adjusted for the effects of intercensal fertility and mortality to provide a residual estimate of intercensal net migration (i.e., treating migration as the balancing item in the fundamental demographic balance equation). Migration is generally concentrated in the age range 20 to 40, ages at which mortality rates are, at least in the absence of HIV/AIDS, relatively low and fertility irrelevant, so residual migration estimates are insensitive to assumptions about fertility and mortality (except in populations severely affected by HIV/AIDS where using these data to estimate migration is not recommended). Such estimates are extremely sensitive, however, to even small changes in census coverage; such errors may be manifest in high age-specific migration rates over age 50, where migration is usually low.

## MIGRATION MEASURES

It is not the purpose of this introduction to provide a comprehensive summary of all the measures and definitions – the interested reader is referred to the UN manual on internal migration (UN Population Division 1970) – but two are of particular importance for the chapters that follow.

**Migration stocks**

Stocks of migrants are typically thought of as numbers of persons (by age group and sex) not born in the civil division of enumeration. The proportions born elsewhere (in the country or in other countries) give a good general sense of the magnitude of in-migration and immigration, but no sense of any dynamic changes that may have occurred recently. However, changes in stocks can be used to estimate immigration (net of any onward or return migration of the foreign-born).

**Migration rates**

Assuming that migration events can be fully and accurately identified, occurrence/exposure rates can be calculated for out-migration or emigration in exactly the same way as for mortality, dividing events in a period by exposure time; such rates can be crude (both sexes, all ages) or age-sex specific. The same is not the case (or at least not usefully) for in-migration or immigration, since the population exposed to the risk of migrating into a civil division is the entire population of the world living elsewhere. In-migration and immigration rates are always calculated by dividing events by the exposure time of the one population group not exposed to risk, the current residents; such rates can be crude (both sexes, all ages) or age-sex specific. Defining rates in this way has the advantage of satisfying the needs of the demographic balancing equation, since rates of gain and loss are measured relative to the same population. This confers a further advantage in that net migration rates can be estimated from the demographic balancing equation as population change between two time points (e.g. censuses) minus gains due to births in the interval plus losses due to deaths in the interval. However, this approach does have the disadvantage of removing the scale limits on "normal"

occurrence/exposure rates; for example, at the extreme, a person moving into a previously unoccupied civil division creates an in-migration rate of infinity.

## DESCRIPTION OF METHODS COVERED

The chapters in this section focus on the estimation and quantitative description of immigration and internal in- and out-migration. They are not meant to provide comprehensive coverage of all measures of migration, and specifically they do not cover the important, but problematic, issue of measuring emigration (other than by mentioning that the method of estimating immigration (net of return/onward migration) of foreigners, can be applied to the data of the main countries of destination of emigrants to get some sense of the age profile and magnitude of emigration.

Chapter 35 concentrates on the basic methods of using data from censuses to estimate the numbers (net of return/onward migration) of immigrants from the change in stock of foreigners, and of internal in- and out-migration from the change in stock by place of birth and from the place of residence at some date prior to the census.

Chapter 36 describes the selection and fitting of a Rogers-Castro multi-exponential model to estimates of migration probabilities (or rates) derived from estimates of the number of migrants/migrations using non-linear optimization procedures.

Chapter 37 describes the multiplicative and log-linear models for capturing, comparing and analysing the mass of inter-regional migration flows from places of origin to places of destination. The chapter also provides an introduction to the method of offsets for extending the use of these models to estimate inter-regional flows from marginal flows (i.e. total flows out of, or into, regions). The intention is to expand the material on the method of offsets into an additional chapter at a later date, which will be placed on the *Tools for Demographic Estimation* website.

## FURTHER READING AND REFERENCES

As mentioned above, *UN Manual VI* (UN Population Division 1970) provides a comprehensive, if dated, introduction to the description and measurement of internal migration. Those looking for an overview of indirect methods of estimating migration are referred to the useful, if also somewhat dated, review by Zaba (1987). More specifically, Hill (1987) attempted to apply the logic underlying the Generalized Growth Balance method of adult mortality estimation (described in Chapter 24) to estimate undocumented migration, while Hill and Queiroz (2010) sought to estimate net migration in parallel with the estimation of mortality, with the focus on improving the mortality estimates. Unfortunately neither method has proved to be particularly successful.

Those interested in reading more about the models of migration (multi-exponential, multiplicative and log-linear) or the method of offsets are referred to work by Rogers, Willekens and colleagues (e.g. Little and Rogers (2007), Raymer and Rogers (2007), Rogers (1980, 1986), Rogers, Little and Raymer (2010) and Willekens (1999)).


Hill K. 1987. "New approaches to the estimation of migration flows from census and administrative data sources", *International Migration Review* **21**(4):1279–1303. doi: 10.2307/2546515

Hill K and B Queiroz. 2010. "Adjusting the general growth balance method for migration", *Revista Brasileira de Estudos de População* **27**(1):7–20. doi: 10.1590/S0102-30982010000100002

Hill K and R Wong. 2005. "Mexico–US migration: Views from both sides of the border", *Population and Development Review* **31**(1):1–18. doi: 10.1111/j.1728-4457.2005.00050.x

Little JS and A Rogers. 2007. "What can the age composition of a population tell us about the age composition of its out-migrants?", *Population, Space and Place* **13**(1):23–19. doi: 10.1002/psp.440

Massey DS, R Alarcon, J Durand and H Gonzalez. 1987. *Return to Aztlan: The Social Process of International Migration from Western Mexico.* Berkeley and Los Angeles: University of California Press.

Raymer J and A Rogers. 2007. "Using age and spacial flow structures in the indirect estimation of migration streams", *Demography* **44**(2):199–223. doi: 10.1353/dem.2007.0016

Rogers A. 1980. "Introduction to multistate mathematical demography", *Environment and Planning A* **12**:489–498. doi: 10.1068/a120489

Rogers A. 1986. "Parameterized multistate population dynamics and projections", *Journal of the American Statistical Association* **81**(393):48–61. doi: 10.1080/01621459.1986.10478237

Rogers A, JS Little and J Raymer. 2010. *The Indirect Estimation of Migration: Methods for Dealing with Irregular, Inadequate, and Missing Data.* Dordrecht: Springer.

UN Population Division. 1967. *Manual IV: Methods for Estimating Basic Demographic Measures from Incomplete Data.* New York: United Nations, Department of Economic and Social Affairs, ST/SOA/Series A/42. http://www.un.org/esa/population/techcoop/DemEst/manual4/manual4.html

UN Population Division. 1970. *Manual VI: Methods of Measuring Internal Migration.* New York: United Nations, Department of Economic and Social Affairs, ST/SOA/Series A/47. http://www.un.org/esa/population/techcoop/IntMig/manual6/manual6.html

UN Population Division. 2011. *International Migration Report 2009: A Global Assessment.* New York: United Nations, Department of Economic and Social Affairs, ST/ESA/Series A/316. http://www.un.org/esa/population/publications/migration/WorldMigrationReport2009.pdf

UN Statistics Division. 2008. *Principles and Recommendations for Population and Housing Censuses v.2.* New York: United Nations, Department of Economic and Social Affairs, ST/ESA/STAT/SER.M/67/Rev2. http://unstats.un.org/unsd/publication/SeriesM/Seriesm_67rev2e.pdf

Willekens FJ. 1999. "Modeling approaches to the indirect estimation of migration flows: From entropy to EM", *Mathematical Population Studies* **7**:239–78. doi: 10.1080/08898489909525459

Zaba B. 1985. *Measurement of Emigration Using Indirect Techniques: Manual for the Collection and Analysis of Data on Residence of Relatives.* Liège: Belgium: Ordina Editions.

Zaba B. 1987. "The indirect estimation of migration: A critical review", *International Migration Review* **21**(4):1395–1445. doi: 10.2307/2546519

# Chapter 35 Estimation of migration from census data

*Rob Dorrington*

## DESCRIPTION OF THE METHODS

Estimating migration from census data is not technically complicated. Provided that the census(es) gather the appropriate information and are reasonably accurate it is possible to produce estimates of net immigration (i.e. immigration less emigration) of the foreign-born population (people born outside a particular country) and internal migration between (to and from) sub-national regions of a country, over the period between two censuses.

To estimate net immigration of foreigners one essentially subtracts from the number of foreign-born people enumerated in a census, the number of foreigners expected to have survived since being enumerated in the previous census.

In a similar way, if the censuses record the sub-national region of birth one can estimate net in-migration (i.e. net in-migration of those born outside the region less net out-migration of those born in the region) between sub-national regions of a country. However, if the census asks of people where they were living at some prior point in time, say at the time of the previous census, one is able to estimate directly the number of surviving migrants (i.e. migrants still alive at the time of the latest census) into and out of each sub-national region of the country since that prior point in time.

In order to estimate the number of migrants from the number of surviving migrants at the time of the second census one needs to add to these figures an estimate of the number of migrants who are expected to have died between moving and the time of the latest census.

If the latest census records other information such as year in which the migrant moved to the place at which the person was counted in the census, it is possible also to establish a trend of migration over time.

Migration is different from fertility and mortality both in that migrating is not final in the sense of a birth or death, but also that we are concerned not only with the population of origin, from which the migrant moved (which corresponds to a population exposed to the risk from which rates of migration akin to those of fertility and mortality can be calculated) but we also have a population to which the migrant moves, the destination population. Apart from this, in order to understand migration one is often interested in distinguishing between different types of migration (whether temporary or more permanent, whether circulatory or unidirectional, etc.). For these reasons there is a much wider range of measures and terminology associated with migration than there is with either fertility or mortality. It is not the purpose of this chapter to cover these issues and the interested reader is referred to the standard texts on the subject such as the *UN Manual VI* (UN Population Division 1970), Shryock and Siegel (1976), Siegel and Swanson (2004).

## DATA REQUIREMENTS AND ASSUMPTIONS

### Tabulations of data required

- To estimate net immigration of foreigners:
  - The number of foreign-born females (males), in five-year age groups, and for an open age interval $A+$, at two points in time, typically two censuses.
  - For the deaths: either a suitable model life table or the numbers of native-born females (males), in five-year age groups, and for an open age interval $A+$, at two points in time, typically two censuses. Failing these, the central crude death rate for the population.
- To estimate sub-national regional net in-migration from place of birth data:
  - The number of females (males) by sub-national region and by sub-national region of birth, in five-year age groups, and for an open age interval $A+$, at two points in time, typically two censuses.
  - For the deaths: either a suitable model life table, the numbers of native-born females (males), in five-year age groups, and for an open age interval $A+$, at two points in time, typically two censuses or numbers of

deaths by region from the vital registration. Failing these, the central crude death rate for the population.
- To estimate internal migration between sub-national regions from place of residence at previous census data:
  - The numbers of females (males) by sub-national region and by sub-national region at some prior date, typically that of the preceding census, in five-year age groups, and for an open age interval *A*+.
- If age-specific numbers are not available, aggregated data is still useful for estimating all-age migration.

**Important assumptions**

- Estimating net immigration of foreigners:
  - Censuses identify all foreign-born people accurately.
  - One is able to estimate the mortality of the foreign-born population accurately (either that the life table used is appropriate, or that the mortality is the same as that implied by the censuses for the native-born (locally-born) national population).
  - No return migration of locally born emigrants.
- Estimating sub-national regional net in-migration from place of birth data:
  - Censuses count the population by sub-national region accurately and identify the region of birth accurately.
  - One is able to estimate the mortality of people moving between two regions accurately (either that the life table used is appropriate, or that the mortality is the same as that implied by the censuses for the native-born national population).
- Estimating internal migration between sub-national regions from data on place of residence at previous census:
  - Latest census identifies correctly all people who have moved from one region to another since the prior date (e.g. previous census).
  - One is able to estimate the mortality of people moving between two regions accurately (either that the life table used is appropriate, or that the mortality is the same as that implied by the censuses for the native-born national population). Since one is estimating in- and out-migration separately (as opposed to net migration) this assumption is of less importance.

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Before applying this method, one should investigate the quality of the data in at least the following dimensions

- age structure of the population (by sub-national region as appropriate); and
- relative completeness of the census counts (by sub-national region as appropriate).

## CAVEATS AND WARNINGS

Estimating migration using place of birth data from two censuses not only requires that the censuses count the population reasonably completely, but that the place of birth be accurately recorded. Often this is not the case, particularly when estimating immigration, where immigrants wish to hide the fact that they are foreign, but also in the case of internal migration where there may have been boundary changes or the respondent is ignorant about the place of birth of the person.

Estimating migration by asking questions of migrants is quite dependent on the census identifying completely all those who have migrated, as well as identifying the place from which moved correctly. To the extent that recent migrants are not yet established as residents of the region to which they have moved at the time of the census, they could be missed in the count.

Net migration, by definition, underestimates the flows of migrants into and out of a region or country. Thus, for example, people who moved into a region and then returned within the period being considered will result in zero net in-migration and yet moved twice.

## APPLICATION OF THE METHOD

### A) Estimating net immigration of foreigners using place of birth data

This method produces estimates of the net immigration of foreigners using place of birth data. It is important to stress that this method does not take into account or measure the immigration of returning native-born people who left the country prior to the previous census and returned before the second census. Thus this method is not recommended for the measurement of immigration where significant return migration of native-born people (for example, after exile or forced migration of refugees) is in progress.

#### *Step 1: Decide on survival factors*

If data on the number of foreign-born people in the population are available by age group for each census then one needs to estimate the survival factors to be applied to the numbers of foreign-born in the first census to estimate the

numbers surviving to the time of the second census. The user can choose between years of life lived factors in five-yearly age groups ($_5L_x$) based on the standard from the General family of United Nations model life tables or one of any of the four families of Princeton model life tables or a model life table of a population experiencing an AIDS epidemic (Timæus 2004) which appear in the *Models* spreadsheet of the associated workbook (see website). This spreadsheet also allows the user to input years of life lived factors in five-yearly age groups of an alternative life table if there is reason to assume that the life table has a similar pattern of mortality to that of the population in question, or failing this, the survival factors can be derived from the proportion of each five-year age group of the native-born population surviving from the first to the second census (assumed to be $n$ years apart, where $n$ is a multiple of 5). Thus $_5S_{s,n}$, $_\infty S_{A-n,n}$ and $S_{B,n}$, the $n$-year survival factor for a group of people aged $x$ to $x+5$ at the previous census, $A-n$ and older at the previous census, and born between censuses, respectively, are estimated as follows:

$$_5S_{x,n} = \frac{_5L_{x+n}}{_5L_x} \text{ or } \frac{_5N^{nb}_{x+n}(t+n)}{_5N^{nb}_{x}(t)}$$

$$_\infty S_{A-n,n} = \frac{T_A}{T_{A-n}} \text{ or } \frac{_\infty N^{nb}_{A}(t+n)}{_\infty N^{nb}_{A-n}(t)}$$

$$\text{and } S_{B,n} = \frac{_nL_0}{nl_0} \text{ or } \frac{_nN^{nb}_{0}(t+n)}{B^{nb}}$$

where the superscript $nb$ represents 'native-born', $_5N^{nb}_x(t)$ represents the native-born population in the census at time $t$ and $B^{nb}$ represents the number of native-born births between time $t$ and $t+n$.

If the data are not available in five-year age groups, the net number of immigrants can still be estimated in total, provided we have an estimate of the crude death rate for the population (which might, in the absence of any evidence to the contrary, be assumed to be that of the native-born population).

***Step 2: Estimate the number of deaths of the immigrants***

If data on the number of foreign-born people in the population are available by age group for two censuses ($n$ years apart) then one needs to estimate the number of deaths of foreign-born people (denoted by the superscript $F$) aged between $x$ and $x+5$ at the first census (at time $t$), $_5D^F_x$, aged $A-n$ and older at the first census, $_\infty D^F_{A-n}$, and those born between the censuses, $D^F_B$, as follows:

$$_5D^F_x = \frac{1}{2}\left(_5N^F_x(t)\cdot {}_5S_{x,n} + {}_5N^F_{x+n}(t+n)\right)\left(\frac{1}{_5S_{x,n}} - 1\right)$$

$$_\infty D^F_{A-n} = \frac{1}{2}\left(_\infty N^F_{A-n}(t)\cdot {}_\infty S_{A-n,n} + {}_\infty N^F_{A}(t+n)\right)\left(\frac{1}{_\infty S_{A-n,n}} - 1\right)$$

$$\text{and } D^F_B = \frac{1}{2}\left(_nN^F_0(t+n)\right)\left(\frac{1}{S_{B,n}} - 1\right)$$

where $_5N^F_x(t)$ represents the number of foreign-born people according to the census at time $t$ who were aged between $x$ and $x+5$.

If data and/or survival factors are not available by age group then one can estimate the total number of deaths of the foreign-born people as follows:

$$_\infty D^F_0 = \frac{n}{2}\left(_\infty N^F_0(t) + {}_\infty N^F_0(t+n)\right)\,{}_\infty m_0$$

where $_\infty m_0$ is an estimate of the crude mortality rate of the population in the country of the census. However, if the age distribution of the foreign-born population is markedly different from that of the population in the country of the census, then this can produce a poor approximation to the true number of deaths.

***Step 3: Estimate the net number of immigrants (of foreigners)***

If data are available by age group for each census then age-specific net immigration can be estimated as follows:

$$\text{Net } {}_5M^F_x = {}_5N^F_{x+n}(t+n) - {}_\infty N^F_x(t) + {}_5D^F_x$$

$$\text{for } x = 0, 5, \ldots, A-5-n$$

where Net $_5M^F_x$ represents the net number of immigrants between times $t$ and $t+n$ who were aged between $x$ and $x+5$ at time $t$. For $x > A-5-n$

$$\text{Net } {}_\infty M^F_{A-n} = {}_\infty N^F_A(t+n) - {}_\infty N^F_{A-n}(t) + {}_\infty D^F_{A-n}.$$

The net number of immigrants of those born between times $t$ and $t+n$ is estimated as follows:

$$\text{Net } M^F_B = {}_nN^F_0(t+n) + D^F_B.$$

If data and/or survival factors are not available by age group then one would estimate of the total net number of immigrants as follows:

$$\text{Net } {}_{\infty}M_0^F = {}_{\infty}N_0^F(t+n) - {}_{\infty}N_0^F(t) + {}_{\infty}D_0^F .$$

**B) Estimating net internal migration between sub-national regions from place of birth data**

Net in-migration into a particular sub-national region from other regions in the country can be estimated in exactly the same way as the international immigration, described above, by replacing the foreign-born population with the population born outside the region.

In addition, applying the same method to data on the change in the numbers of population born in (rather than outside) and living outside the region of interest allows us to estimate the net out-migration of those born in the region to other regions in the country. Subtracting this from the net in-migration of those born outside the region gives an estimate of the overall net in-migration into the region of interest.

If there is reason to suspect that there is a material difference in the mortality experienced by those born outside who moved into the region and those born in the region who moved out, and one has appropriate survival factors then one could apply different survival factors to each when estimating the net number of migrants. However, in practice it is likely that inaccuracies in the census data on place of residence at previous census are likely to outweigh any increase in accuracy achieved by using differential mortality.

**C) Estimating internal migration between sub-national regions from place of residence at previous survey**

Net sub-national inter-regional migration is estimated directly from the numbers of people in each region at the time of the census who moved since the previous census by place (e.g. region) they were in at a given prior date (e.g. at the time of the previous census). Confining the estimates to inter-regional flows the sum of the numbers of inter-regional in-migrants should be equal to the sum of inter-regional out-migrants; however, if the data include immigration to the sub-national regions from outside the country one can extend the estimates of in-migration to include international immigration into each region.

Since one of the major areas of interest is the magnitude of inter-regional flows of the population, one is as interested in the total numbers of migrants between regions as one is in the age distributions of particular flows.

The number of migrants is derived from the number of surviving in- and out-migrants as follows:

$${}_5M_x = \left( {}_5I'_x - {}_5O'_x + \left( {}_5I'_x - {}_5O'_x \right)_x \Big/ {}_5S_x \right) \Big/ 2$$

where the superscript (′) designates numbers surviving and ${}_5I'_x$ and ${}_5O'_x$ respectively represent the number of surviving in-migrants into, and the number of surviving out-migrants from, a particular region at the time of the second census who were aged between $x$ and $x+5$ at the second census.

**WORKED EXAMPLE**

This example uses data on the numbers of males in the population from the South African Census in 2001 and a 'census replacement survey', the Community Survey in 2007. (Although the survey was conducted approximately 5.35 years after the night of the census in 2001, it is assumed for the purposes of presentation here to have been exactly five years after the census in 2001.) The examples appear in the *Migration_South Africa_males.xlsx* workbook (see website).

**A) Estimating net immigration of foreigners using place of birth**

***Step 1: Decide on survival factors***

The survival factors are shown in the fifth column of Table 35.1. The values are derived from (the years of life lived in each age group of) the alternative life table entered in the *Models* spreadsheet, for those aged 20 to 24 last birthday and those aged 80 and over at the time of the first census, and those born between the two censuses, as follows:

$${}_5S_{20,5} = \frac{{}_5L_{25}}{{}_5L_{20}} = \frac{4.3382}{4.4975} = 0.96458$$

$${}_{\infty}S_{80,5} = \frac{T_{85}}{T_{80}} = \frac{0.75180}{1.19603} = 0.40912$$

$$\text{and } S_{B,5} = \frac{{}_5L_0}{5l_0} = \frac{4.707549}{5} = 0.94151 .$$

***Step 2: Estimate the number of deaths***

Since we have data on the number of foreign-born people in the population by age group for each census we can estimate the number of deaths of foreign-born people which occurred in the period between the two censuses by age group using

the numbers of foreigners in each census given in the second and third columns of Table 35.1. For those aged 20 to 24 last birthday and those aged 80 and over at the time of the first census, and those born between the two censuses, the calculations are as follows:

$$
\begin{aligned}
{}_5D_{20}^F &= \frac{1}{2}\left({}_5N_{20}^F(2001)\cdot {}_5S_{20,5} + {}_5N_{25}^F(2006)\right)\left(\frac{1}{{}_5S_{20,5}} - 1\right) \\
&= \left(69787 \cdot 0.96458 + 95763\right)\left(\frac{1}{0.96458} - 1\right) \\
&= 2994
\end{aligned}
$$

$$
\begin{aligned}
{}_\infty D_{80}^F &= \frac{1}{2}\left({}_\infty N_{80}^F(2001)\cdot {}_\infty S_{80,5} + {}_\infty N_{85}^F(2006)\right)\left(\frac{1}{{}_\infty S_{80,5}} - 1\right) \\
&= \left(\left(7658 + 4455\right)0.40912 + 5305\right)\left(\frac{1}{0.40912} - 1\right) \\
&= 7410
\end{aligned}
$$

and

$$
\begin{aligned}
D_B^F &= \frac{1}{2}\left({}_5N_0^F(2006)\right)\left(\frac{1}{S_{B,5}} - 1\right) = 12577\left(\frac{1}{0.94151} - 1\right) \\
&= 391.
\end{aligned}
$$

If data and/or survival factors were not available by age group then one could estimate the total number of deaths of the foreign born people as follows, given an estimate of the crude mortality rate in the population of 14 per 1,000:

$$
\begin{aligned}
{}_\infty D_0^F &= \frac{5}{2}\left({}_\infty N_0^F(2001) + {}_\infty N_0^F(2006)\right){}_\infty m_0 \\
&= \frac{5}{2}\left(611423 + 754608\right)\frac{14}{1000} = 47811.
\end{aligned}
$$

***Step 3: Estimate the net number of immigrants (of foreigners)***

Since data are available by age group for each census, age-specific net immigration of those born outside the country can be estimated as follows:

*TABLE 35.1 Estimation of deaths of foreign-born and the net number of immigrants by age group, South Africa, 2001–2006*

| Age | 2001 | 2006 | $x$ | ${}_5S_x$ | Age at 2nd census | $D_F$ | Net $M$ |
|---|---|---|---|---|---|---|---|
| | | | $B$ | 0.94151 | | | |
| 0–4 | 8,963 | 12,577 | 0 | 0.97896 | 0–4 | 391 | 12,968 |
| 5–9 | 10,390 | 13,724 | 5 | 0.99547 | 5–9 | 242 | 5,003 |
| 10–14 | 13,508 | 13,998 | 10 | 0.99427 | 10–14 | 55 | 3,664 |
| 15–19 | 27,835 | 27,943 | 15 | 0.98602 | 15–19 | 119 | 14,555 |
| 20–24 | 69,787 | 59,493 | 20 | 0.96458 | 20–24 | 616 | 32,275 |
| 25–29 | 87,381 | 95,763 | 25 | 0.93161 | 25–29 | 2,994 | 28,970 |
| 30–34 | 73,338 | 100,450 | 30 | 0.90960 | 30–34 | 6,675 | 19,743 |
| 35–39 | 66,663 | 85,490 | 35 | 0.89780 | 35–39 | 7,563 | 19,715 |
| 40–44 | 59,152 | 75,684 | 40 | 0.89092 | 40–44 | 7,701 | 16,721 |
| 45–49 | 45,184 | 66,113 | 45 | 0.88633 | 45–49 | 7,274 | 14,234 |
| 50–54 | 40,398 | 55,913 | 50 | 0.87224 | 50–54 | 6,154 | 16,883 |
| 55–59 | 30,640 | 42,833 | 55 | 0.84731 | 55–59 | 5,717 | 8,153 |
| 60–64 | 24,376 | 34,433 | 60 | 0.80885 | 60–64 | 5,442 | 9,234 |
| 65–69 | 17,895 | 25,588 | 65 | 0.75468 | 65–69 | 5,353 | 6,564 |
| 70–74 | 13,561 | 18,989 | 70 | 0.66991 | 70–74 | 5,281 | 6,375 |
| 75–79 | 10,238 | 12,850 | 75 | 0.56388 | 75–79 | 5,404 | 4,693 |
| 80–84 | 7,658 | 7,461 | 80+ | 0.40912 | 80–84 | 5,118 | 2,341 |
| 85+ | 4,455 | 5,305 | | | 85+ | 7,410 | 602 |
| **TOTAL** | **611,423** | **754,608** | | | **TOTAL** | **79,509** | **222,693** |

$$\text{Net } {}_5M_{20}^F = {}_5N_{25}^F(2006) - {}_\infty N_{20}^F(2001) + {}_5D_{20}^F$$
$$= 95763 - 69787 + 2994 = 28970$$

$$\text{Net } {}_\infty M_{80}^F = {}_\infty N_{85}^F(2006) - {}_\infty N_{80}^F(2001) + {}_\infty D_{80}^F$$
$$= 5305 - (7658 + 4455) + 7410 = 602$$

$$\text{Net } M_B^F = {}_5N_0^F(2006) + D_B^F = 12577 + 391 = 12968.$$

If data and/or survival factors were not available by age group then one could estimate the total net number of immigrants as follows:

$$\text{Net } {}_\infty M_0^F = {}_\infty N_0^F(2006) - {}_\infty N_0^F(2001) + {}_\infty D_0^F$$
$$= 754608 - 611423 + 47811 = 190996.$$

**B) Estimating sub-national regional net in-migration using place of birth**

The second and third column of Table 35.2 show the numbers of people living in the Western Cape province of South Africa who were born outside the province, as counted by the 2001 Census and the 2007 Community Survey, respectively. Although the same survival factors (column 5) have been used as were used in the example of Method A, this should not be the case if it was thought that the mortality experience of native-born and immigrants were very different. The final column of Table 35.2 gives the net numbers of migrants into the Western Cape who were born in provinces other than the Western Cape for the different age groups. Thus in total 213,911 people born outside the Western Cape moved to the Western Cape (after excluding those who moved out).

The second and third columns of Table 35.3 present the numbers of people living in provinces other than the Western Cape who were born in the Western Cape, as

*TABLE 35.2 Estimation of the net number of in-migrants of those born outside by age group, Western Cape, South Africa, 2001–2006*

| Age | 2001 | 2006 | $x$ | ${}_5S_x$ | Age at 2nd census | $D_O$ | Net $M$ (born out) |
|---|---|---|---|---|---|---|---|
| | | | $B$ | 0.94151 | | | |
| 0–4 | 16,443 | 19,012 | 0 | 0.97896 | 0–4 | 591 | 19,602 |
| 5–9 | 24,406 | 28,743 | 5 | 0.99547 | 5–9 | 482 | 12,782 |
| 10–14 | 31,134 | 30,792 | 10 | 0.99427 | 10–14 | 125 | 6,511 |
| 15–19 | 44,478 | 53,933 | 15 | 0.98602 | 15–19 | 245 | 23,043 |
| 20–24 | 74,011 | 82,526 | 20 | 0.96458 | 20–24 | 896 | 38,944 |
| 25–29 | 80,187 | 89,522 | 25 | 0.93161 | 25–29 | 2,954 | 18,466 |
| 30–34 | 65,833 | 90,783 | 30 | 0.90960 | 30–34 | 6,074 | 16,670 |
| 35–39 | 56,393 | 76,475 | 35 | 0.89780 | 35–39 | 6,776 | 17,417 |
| 40–44 | 44,420 | 59,692 | 40 | 0.89092 | 40–44 | 6,268 | 9,567 |
| 45–49 | 32,862 | 47,612 | 45 | 0.88633 | 45–49 | 5,338 | 8,529 |
| 50–54 | 28,178 | 37,969 | 50 | 0.87224 | 50–54 | 4,303 | 9,409 |
| 55–59 | 1 9,983 | 30,205 | 55 | 0.84731 | 55–59 | 4,012 | 6,039 |
| 60–64 | 17,569 | 25,593 | 60 | 0.80885 | 60–64 | 3,832 | 9,442 |
| 65–69 | 11,216 | 20,802 | 65 | 0.75468 | 65–69 | 4,137 | 7,371 |
| 70–74 | 8,365 | 12,612 | 70 | 0.66991 | 70–74 | 3,426 | 4,822 |
| 75–79 | 5,919 | 8,434 | 75 | 0.56388 | 75–79 | 3,458 | 3,528 |
| 80–84 | 4,063 | 5,061 | 80+ | 0.40912 | 80–84 | 3,248 | 2,390 |
| 85+ | 2,152 | 2,183 | | | 85+ | 3,413 | –620 |
| **TOTAL** | **567,613** | **721,949** | | | **TOTAL** | **59,576** | **213,911** |

counted by the 2001 Census and the 2007 Community Survey, respectively. The net number of out-migrants of those born in the Western Cape (i.e. the number of people born in the Western Cape who moved out, less those who have returned) is given in column 8. The negative numbers mean that there was negative net out-migration (i.e. the number of those born in the Western Cape who moved to other provinces in the period was less than the number born in the Western Cape who were living outside who returned during the period). Thus the total of –19,017 means that the number of people born in the Western Cape, who returned to the Western Cape during the period having lived in another province until 2001 exceed those who were born in the Western Cape and moved to another province in the period by 19,017.

These estimates were derived using the same survival factors as were used for those born outside the Western Cape who moved into the province, but if there was reason to suppose that the mortality differed for those born in the Western Cape who moved out, then a different set of survival factors would be used to estimate the Net *M* (born in) numbers.

The overall net in-migration for the province is thus given in the final column of Table 35.3. Thus in total 232,928 more people moved into the Western Cape than left the Western Cape to live in another province.

In this example those born outside the province include those born outside the country and thus the overall net migration includes immigrants who settle in the province. Excluding the foreign-born from Table 35.2 would produce numbers of internal in-migrants net of internal out-migrants, and the sum of these numbers for all the provinces together would be zero.

### C) Estimating internal migration between sub-national regions from data on place of residence at previous census

Table 35.4 presents the results of the answers to the question about place (province in this example) of residence at the

*TABLE 35.3 Estimation of the net number of out-migrants of those born inside by age group, Western Cape, South Africa, 2001–2006*

| Age | 2001 | 2006 | $x$ | ${}_5S_x$ | Age at 2nd census | $D_I$ | Net *M* (born in) | Overall Net *M* |
|---|---|---|---|---|---|---|---|---|
| | | | *B* | 0.94151 | | | | |
| 0–4 | 22,055 | 11,747 | 0 | 0.97896 | 0–4 | 365 | 12,112 | 7,490 |
| 5–9 | 21,895 | 12,509 | 5 | 0.99547 | 5–9 | 367 | –9,180 | 21,962 |
| 10–14 | 21,382 | 11,593 | 10 | 0.99427 | 10–14 | 76 | –10,226 | 16,737 |
| 15–19 | 18,265 | 13,455 | 15 | 0.98602 | 15–19 | 100 | –7,827 | 30,870 |
| 20–24 | 14,645 | 10,477 | 20 | 0.96458 | 20–24 | 202 | –7,587 | 46,531 |
| 25–29 | 13,501 | 9,534 | 25 | 0.93161 | 25–29 | 434 | –4,676 | 23,142 |
| 30–34 | 13,118 | 11,047 | 30 | 0.90960 | 30–34 | 867 | –1,587 | 18,257 |
| 35–39 | 12,121 | 14,614 | 35 | 0.89780 | 35–39 | 1,319 | 2,815 | 14,602 |
| 40–44 | 11,725 | 12,195 | 40 | 0.89092 | 40–44 | 1,311 | 1,384 | 8,183 |
| 45–49 | 10,335 | 10,538 | 45 | 0.88633 | 45–49 | 1,285 | 98 | 8,431 |
| 50–54 | 9,211 | 9,881 | 50 | 0.87224 | 50–54 | 1,221 | 768 | 8,642 |
| 55–59 | 7,264 | 10,568 | 55 | 0.84731 | 55–59 | 1,362 | 2,720 | 3,319 |
| 60–64 | 6,691 | 7,723 | 60 | 0.80885 | 60–64 | 1,250 | 1,710 | 7,732 |
| 65–69 | 4,643 | 5,297 | 65 | 0.75468 | 65–69 | 1,265 | –128 | 7,499 |
| 70–74 | 3,954 | 3,766 | 70 | 0.66991 | 70–74 | 1,182 | 304 | 4,517 |
| 75–79 | 2,331 | 2,384 | 75 | 0.56388 | 75–79 | 1,240 | –330 | 3,858 |
| 80–84 | 1,402 | 2,140 | 80+ | 0.40912 | 80–84 | 1,336 | 1,145 | 1,244 |
| 85+ | 707 | 555 | | | 85+ | 1,024 | –531 | –89 |
| **TOTAL** | **195,246** | **160,023** | | | **TOTAL** | **16,206** | **–19,017** | **232,928** |

time of the 2001 Census given by those counted in each of the provinces in the 2007 Community Survey. (In actual fact the question asked whether the person was staying at the same place at the time of the prior census and if not, where they were staying at the time they moved to the place at which they were counted in the Community Survey. However, work by Dorrington and Moultrie (2009) shows that using these data and the year of movement to back-project the population in order to estimate the numbers by province of residence at the time of the previous survey suggests that the assumption that there was only one move in the five years since the previous census was reasonably accurate.)

By far the largest numbers of migrants are those that moved within each of the provinces, however, these have been excluded from Table 35.4 because one is usually more interested in interprovincial migration than migration within a province.

In addition to the all-age numbers in Table 35.4 (in actual fact these numbers exclude, as is often the case, migration of those born between the census and survey) one can also produce numbers of in- and out-migration by age groups as shown in Table 35.5. For completeness these numbers include estimates of the number of migrants who were born since the previous census. However, relative to the number of migrants at other ages, these numbers look implausibly high, and the reason for this is discussed below.

The net number of migrants is estimated for those aged 25–29 at the time of the Community Survey (i.e. were aged 20–24 at the time of the 2001 census), for example, as follows:

$$
\begin{aligned}
{}_5M_x &= \left(20675 - 5649 + \left(20675 - 5649\right)/0.96458\right)/2 \\
&= 15301.
\end{aligned}
$$

## DIAGNOSTICS, ANALYSIS AND INTERPRETATION

### Checks and validation

Perhaps the simplest check, on the reasonableness of the 'shape' (i.e. distribution of the numbers by age) of the estimates but not the level, is to see if it conforms to the standard shape (or a variation thereof). Rogers and Castro (1981a; 1981b) point out that the distribution of the number (or rate) of in- and out-migrants tends to conform to standard patterns, with a peak in the young adult ages (usually associated with seeking employment), a second, usually less pronounced peak amongst very young

*TABLE 35.4 Interprovincial migration, South Africa, 2001–2006*

| | Province where counted (destination) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Previous residence (origin)** | **WC** | **EC** | **NC** | **FS** | **KZ** | **NW** | **GT** | **MP** | **LM** | **Total** |
| WC | | 12,173 | 4,060 | 1,745 | 3,221 | 2,113 | 16,400 | 1,405 | 874 | 41,992 |
| EC | 52,239 | | 1,120 | 7,187 | 25,209 | 14,430 | 28,633 | 4,693 | 2,116 | 135,626 |
| NC | 4,813 | 1,942 | | 3,480 | 908 | 3,728 | 4,956 | 1,062 | 357 | 21,246 |
| FS | 2,943 | 3,145 | 2,546 | | 2,352 | 12,733 | 19,920 | 4,293 | 1,963 | 49,896 |
| KZ | 6,762 | 7,015 | 631 | 2,358 | | 3,573 | 50,980 | 8,886 | 1,194 | 81,399 |
| NW | 1,478 | 907 | 9,811 | 5,555 | 2,329 | | 47,633 | 3,090 | 4,337 | 75,140 |
| GT | 24,891 | 12,948 | 3,962 | 11,437 | 18,145 | 32,433 | | 18,598 | 15,133 | 137,547 |
| MP | 2,134 | 1,317 | 280 | 1,724 | 4,546 | 5,767 | 42,941 | | 8,628 | 67,338 |
| LM | 2,754 | 1,583 | 255 | 1,709 | 2,209 | 9,773 | 81,394 | 24,211 | | 123,889 |
| OSA | 21,221 | 5,467 | 1,209 | 9,584 | 10,933 | 11,437 | 51,873 | 8,335 | 9,286 | 129,346 |
| DNK | 500 | 3 | 15 | 124 | 132 | 78 | 228 | 89 | 0 | 1,170 |
| UNS | 1,058 | 1,029 | 107 | 208 | 875 | 508 | 3,558 | 408 | 633 | 8,384 |
| **TOTAL** | **120,794** | **47,528** | **23,996** | **45,111** | **70,860** | **96,573** | **348,516** | **75,070** | **44,524** | **872,973** |

WC = Western Cape, EC = Eastern Cape, NC = Northern Cape, FS = Free State, KZN = KwaZulu-Natal, NW = North West, GT = Gauteng, MP = Mpumulanga, LM = Limpopo, OSA = Outside SA, DNT = Do not know, UNS = Unspecified

children falling to a trough amongst young teenagers (the size depending on the extent to which it is families rather than individuals moving in the young to middle aged adults). Sometimes there is also a 'hump' (or trough) around retirement age if there is a strong flow of migrants moving to (or away from) the place to retire.

These patterns (not necessarily the same pattern) apply to in- and out-migration flows separately, but not necessarily to net migration (which is the difference between the two flows) unless one flow (either the in-migration or the out-migration) is much greater than the other.

Figure 35.1 illustrates this using some of the estimates calculated above, expressed as proportions of the total number in each case (to allow them to be presented on a single figure). From this we can see that in broad terms (with the exception in some cases, where the proportion of migrants at the very young ages looks implausibly high) each conforms to the expected shape.

The net out-migrants of those born in the Western Cape (excluded from the figure for ease of illustration) does not conform to a standard model of migration, which could indicate these numbers are not very reliable, however, they are small relative to the in-migration of those born outside the province, and thus such a deviation may tolerated. In addition to this there are two other features to be noted from Figure 35.1. The first is that the out-migration from the Western Cape as estimated from data on place of residence at previous census, suggests that adult out-migrants peak at a somewhat older age (and possibly are likely to represent family rather than individual migration). The second is the fact that the net immigration into the country follows the standard shape which indicates that the flow into the country is much stronger than the return flow of those migrants.

If the census asked place of birth and place of residence at the previous census then one can compare the two estimates of net in-migration into a specific sub-national region. If they are similar this gives one some confidence in the results. In the case of the place of birth data for South Africa the net number of in-migrants into the Western Cape is 232,928 (Table 35.3) while the estimate from the data on place of residence at the time of the previous census data produced an estimate of 92,194 (Table 35.4), which suggests that one or both of these sets of data are suspect.

The most basic check of the estimates of migration is to

*TABLE 35.5 Estimation of the net number of in-migrants by age group, Western Cape, South Africa, 2001–2006*

| Age | Surviving in-migrants ($I'$) | Surviving out-migrants ($O'$) | $x$ | ${}_5S_x$ | Net in-migrants |
|---|---|---|---|---|---|
| 0–4 | 20,846 | 11,747 | $B$ | 0.94151 | 9,381 |
| 5–9 | 6586 | 3,554 | 0 | 0.97896 | 3,065 |
| 10–14 | 6685 | 2,882 | 5 | 0.99547 | 3,812 |
| 15–19 | 10402 | 3,967 | 10 | 0.99427 | 6,454 |
| 20–24 | 21266 | 4,488 | 15 | 0.98602 | 16,897 |
| 25–29 | 20675 | 5,649 | 20 | 0.96458 | 15,301 |
| 30–34 | 15584 | 6,008 | 25 | 0.93161 | 9,928 |
| 35–39 | 10584 | 5,098 | 30 | 0.90960 | 5,758 |
| 40–44 | 7264 | 3,045 | 35 | 0.89780 | 4,458 |
| 45–49 | 4648 | 2,714 | 40 | 0.89092 | 2,053 |
| 50–54 | 3095 | 1,500 | 45 | 0.88633 | 1,698 |
| 55–59 | 3940 | 935 | 50 | 0.87224 | 3,225 |
| 60–64 | 3776 | 527 | 55 | 0.84731 | 3,541 |
| 65–69 | 3127 | 818 | 60 | 0.80885 | 2,582 |
| 70–74 | 1540 | 437 | 65 | 0.75468 | 1,282 |
| 75–79 | 561 | 206 | 70 | 0.66991 | 442 |
| 80–84 | 797 | 116 | 75 | 0.56388 | 944 |
| 85+ | 264 | 47 | 80+ | 0.40912 | 374 |
| **TOTAL** | **141,640** | **53,739** | | | **91,194** |

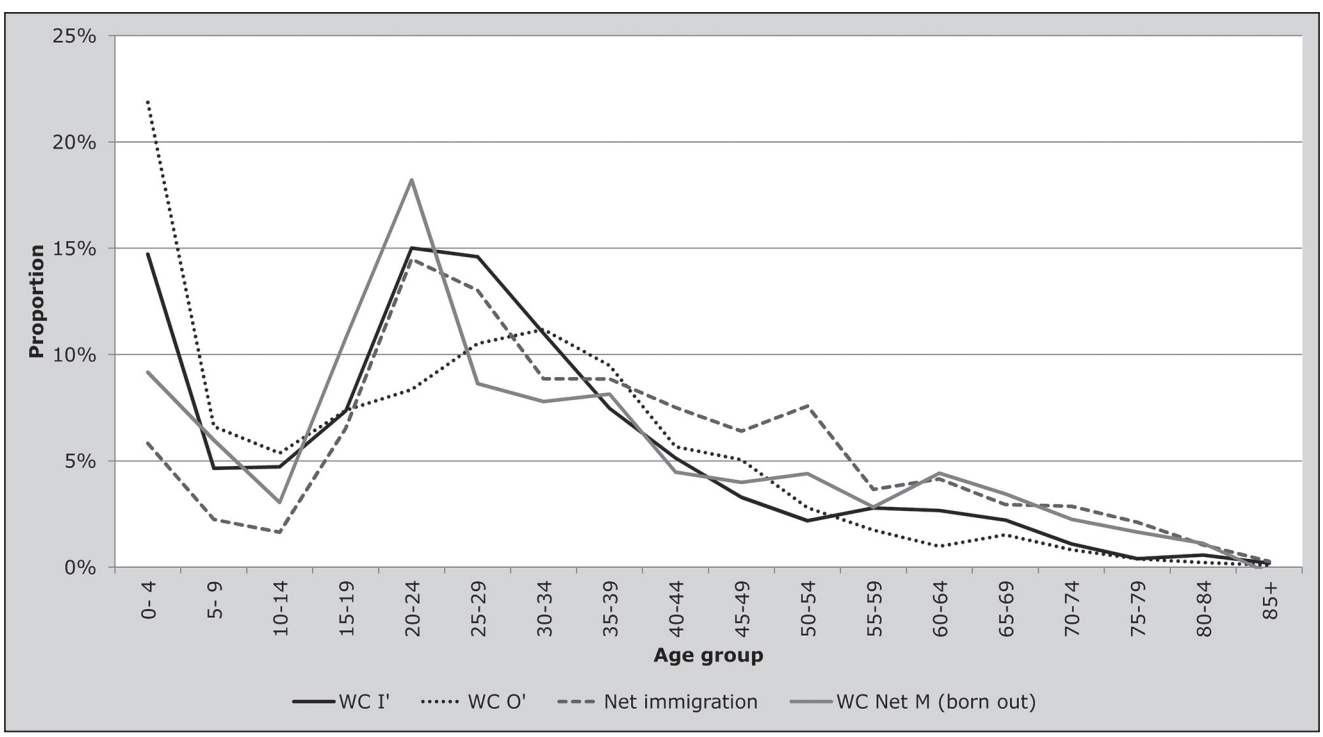
project the population (of the country or the province) at the first census to the time of the second census making use of the estimates of the number of migrants and compare that with the census estimates from the second, more recent, census to see how well the two match, especially in the age range in which migration is concentrated. In the case of the net in-migration into the Western Cape, projecting the population forward from 2001 using the estimates derived from the change in the numbers by place of birth produced a much closer fit to the population in the 20–29 year age range, suggesting that the data on place of birth are probably more complete than those on the place of residence at the date of the previous census. To some extent this is supported by a comparison of the change in the number of foreign-born in the country between the two censuses, 222,693 (Table 35.1) with the sum of the numbers who reported that they had moved from outside South Africa to one of the provinces since the previous census, 129,346 (Table 35.4).

Ideally, if one had independent estimates of the number of migrants one might compare those numbers against estimates using the above methods. Unfortunately, reliable independent estimates are rare. Although most countries try to record people entering and leaving the country, these data are often not reliable, particularly in developing countries with relative porous borders. And unless the country is extremely well regulated and maintains a complete and accurate register of the population, the only other way to measure internal migration is through migration-specific surveys, which tend to be much more useful for understanding the type of migration (whether permanent, temporary, cyclical, etc.) than for producing reliable estimates of the number of migrants, given the often less structured situation that (particularly recent) migrants find themselves living in and an understandable reluctance to identify themselves as being migrants.

**Interpretation**

Considering the numbers of migrants estimated from the data on place of residence at the previous census given in Table 35.4 (and taking into account the suspicion that these probably underestimate the true migration), some 2–4 per cent of the population changed province of residence in

*FIGURE 35.1 Age distribution of selected migrant flows, South African males, 2001–2006*



the 5 years between the 2001 Census and the Community Survey. Had we included the number who moved within, but did not change, province then between 7 and 15 per cent of the population moved in the 5-year period.

The main provinces of destination are Gauteng (by a big margin) and Western Cape, which are predominantly urban and the wealthiest provinces. The main provinces of origin are Gauteng (inspection of the age distribution would show that this is mainly return migration of 'retiring' workers) Eastern Cape and Limpopo, which are poor, mainly rural provinces, from which people seeking work migrate to the urban areas.

It appears that migration is predominantly of individuals (seeking work) rather than of families.

## METHOD-SPECIFIC ISSUES WITH INTERPRETATION

### Scanning errors

A particular feature of the data relying on province of birth is the apparently relatively high number of children born since the first census who have moved to another province. In all likelihood this is an artefact of the data capturing process. Scanning was used to capture the data from the questionnaires on which Western Cape was coded as a "1", written in the appropriate space by hand. It appears that in a small percentage of cases the scanner might have had trouble distinguishing a handwritten "1" from a handwritten "7" (the code for Gauteng). The result of this is, for example, that some of the children coded as having been born outside the province in which they were counted, and thus appear to be migrants, but probably were not. Even though the percentage error in scanning is very small, the number of births can be large relative to the number migrants, and thus the error can produce noticeable errors. Since an increasing number of developing countries are using scanning to capture data, this sort of problem may be quite common.

Where scanning errors or other situations make it impossible to produce reliable estimates of the number of migrants of those born since the previous census one can use CWR from second census as follows:

$$\text{Net}_5M_0 = \frac{1}{4}CWR_0 \cdot \text{Net}\ {}_{30}M_{15}^{f}$$

for those born in the most recent five years, and

$$\text{Net}_5M_5 = \frac{3}{4}CWR_5 \cdot \text{Net}\ {}_{30}M_{20}^{f}$$

for those born in the five years before that if the censuses are 10 years apart, where $CWR_x$ represents the ratio of the number of children aged between $x$ and $x+5$ to the number of women in the population aged between $15+x$ and $45+x$ in the population (regional or national) at the time of the second census, and ${}_{30}M_x^f$ represents the number of women migrants aged between $x$ and $x+30$.

Applying this to the data for the Western Cape suggest that the number of migrants born since the previous census should be less than half the numbers being estimated from the data on place of birth.

## DETAILED DESCRIPTION OF METHOD

### Mathematical exposition

The indirect estimation of migration derives from the balance equation for two censuses $n$ years apart, namely:

$$\begin{aligned} {}_5N_{x+n}(t+n) &= {}_5N_x(t) - {}_5D_x + {}_5I'_x - {}_5O'_x \\ &= {}_5N_x(t) - {}_5D_x + {}_5M'_x \end{aligned}$$

where ${}_5M'_x = {}_5I'_x - {}_5O'_x$ is the net (i.e. in less out) number of in-migrants, aged $x$ to $x+5$ at the time of the first census, surviving to the second census, and ${}_5D_x$, ${}_5I'_x$ and ${}_5O'_x$, represent the number of deaths, surviving in-migrants and out-migrants, aged $x$ to $x+5$ at the time of the first census, who died or moved in the period between the censuses.

For those born after the first census the equation becomes:

$${}_nN_0(t+n) = B - D_B + M'_B$$

and those in the open age interval:

$${}_\infty N_A(t+n) = {}_\infty N_{A-n}(t) - {}_\infty D_{A-n} + {}_\infty M'_{A-n}$$

where $B$ represents the number of births in the population between the two censuses, $D_B$ the number of deaths of those births in the period between the censuses and $M'_B$ the net number of surviving migrants, born outside the country in the period between the two censuses, ${}_\infty D_{A-n}$ the number of deaths in the intercensal period aged $A-n$ and older at the time of the first census, and ${}_\infty M'_{A-n}$ the net number of migrants aged $A-n$ and older at the time of the first census. Thus,

$${}_5M'_x = {}_5N_{x+n}(t+n) - {}_5N_x(t) + {}_5D_x$$

$$M'_B = {}_nN_0(t+n) - B + D_B$$

$${}_\infty M'_{A-n} = {}_\infty N_A(t+n) - {}_\infty N_{A-n}(t) + {}_\infty D_{A-n}$$

or alternatively

$${}_5M'_x = {}_5N_{x+n}(t+n) - {}_5N_x(t)\,{}_5S_x$$

$$M'_B = {}_nN_0(t+n) - BS_B$$

$${}_\infty M'_{A-n} = {}_\infty N_A(t+n) - {}_\infty N_{A-n}(t)\,{}_\infty S_{A-n}$$

where ${}_5S_x$, $S_B$ and ${}_\infty S_{A-n}$ represent the proportion of the populations aged $x$ to $x+5$ at the time of the first census, born between the censuses, and aged $A-n$ and older at the time of the first census, respectively, surviving to the second census.

The net number of migrants can thus be estimated from the net number surviving to the second census as follows:

$${}_5M_x = \left({}_5M'_x + {}_5M'_x / {}_5S_x\right)\Big/ 2 = {}_5M'_x \frac{\left({}_5S_x + 1\right)}{2\,{}_5S_x}$$

$$M_B = M'_B \frac{\left(S_B + 1\right)}{2S_B}$$

$${}_\infty M_{A-n} = {}_\infty M'_{A-n} \frac{\left({}_\infty S_{A-n} + 1\right)}{2\,{}_\infty S_{A-n}}\,.$$

Unfortunately, since the net number of migrants is usually small relative to the size of the population, age misstatement or errors in either or both census counts can lead to very poor estimates being produced. Better estimates of the net number of immigrants into a country can be produced by confining one's attention to the population of foreigners (defined as those born outside the country) and assuming that return migration of emigrants from the country of interest is insignificant. Thus one replaces each of the symbols above by equivalents specific to the foreign-born population in the country. Since it is unlikely that one has an accurate record of the number of the foreign-born deaths these need to be estimated in one of the following ways:

- *Option 1 (Life table survival ratios)*: Applying rates from a suitable model life table, then

$${}_5S_x = \frac{{}_5L_{x+n}}{{}_5L_x},\ S_B = \frac{{}_nL_0}{n \cdot l_0} \text{ and } {}_\infty S_{A-n} = \frac{T_A}{T_{A-n}}$$

- *Option 2 (Census survival ratios)*: Assuming that emigration of the native-born population is insignificant and that the proportions surviving are the same as those in the native-born population, then

$${}_5S_x = \frac{{}_5N^{nb}_{x+n}(t+n)}{{}_5N^{nb}_x(t)},\ S_B = \frac{{}_nN^{nb}_0}{B^{nb}} \text{ and}$$

$${}_\infty S_{A-n} = \frac{{}_\infty N^{nb}_A(t+n)}{{}_\infty N^{nb}_{A-n}(t)}$$

where the superscript *nb* designates native-born.

- *Option 3 (Vital registration)*: Where one has access to numbers of births and deaths from another source such as vital registration (which is only likely to be the case, if at all, with internal migration), one could work with deaths and births corresponding to the migrant population directly instead of survival ratios to estimate the net number of surviving in-migrants. Alternatively the net number of migrants can be derived as above by setting

$${}_5S_x = 1 - \frac{{}_5D_x}{{}_5N_x(t)},\ S_B = \frac{D_B}{B} \text{ and } {}_\infty S_{A-n} = \frac{D_{A-n}}{{}_\infty N_{A-n}(t)}$$

where the births and deaths are from the vital registration.

However, for most developing countries, particularly those in Africa, vital registration systems are too incomplete to be used in this way.

**Internal migration**

When it comes to internal migration one can estimate net in-migration (i.e. in-migration of those born outside the region less out-migration of those born outside the region who had previously moved into the region) into each sub-national region of those born outside the region by making use of place of birth information to identify the change in numbers of those born outside the region, in the same way as described above. However, since one also has the place of residence of those born in the region who have moved out of the region since birth (but not emigrated) one can also estimate the net out-migration of those born in the region (i.e. out-migration of those born in the region less those born in the region who have returned after having previously moved out of the region) by applying the method described above to the population born in the region (as opposed to those born outside the region).

When estimating the survival of those born in the various regions the census survival ratios could have an advantage

over the life table survival ratios in that any under or over count of the population by region, may well be matched by a similar distortion in the national population and hence in the survival ratios, thus resulting in a more accurate estimate of the number of migrants than would be produced by using life table survival ratios.

Apart from place of birth, a census can ask of those who moved since the previous census (or some other suitable date) where they were at that census (or some other suitable date) which allows one to measure out-migration and hence (gross) in-migration separately for each sub-national region.

If the census asks for the year when the migrant moved (or how long the person has been living in the place where counted in the second census) one can get a sense of the timing of migration, and estimate yearly migration rates. This is a complicated process and is not covered here, but the interested reader is referred to the paper by Dorrington and Moultrie (2009).

**Working with total numbers only**

If age-specific numbers are not available or the allocation to age is considered to be unreliable one can still produce estimates by age by estimating the total number of migrants as described below, and then apportioning this total to the age groups using either an age distribution for the same population at a different time (since the age distribution of migration flows tend be consistent over time, or (more likely) an appropriate standard model, such as that described by Rogers and Castro (1981a; 1981b). Accordingly,

$$\text{Net } {}_{\infty}M_0^F = {}_{\infty}N_0^F(t+n) - {}_{\infty}N_0^F(t) + {}_{\infty}D_0^F$$

where ${}_{\infty}D_0^F = \frac{n}{2}\left( {}_{\infty}N_0^F(t) + {}_{\infty}N_0^F(t+n)\right) {}_{\infty}m_0$

and ${}_{\infty}m_0$ is an estimate of the crude mortality rate of the population in the country of the census.

**Limitations**

The primary limitation of using censuses to estimate immigration and net in-migration is the quality of the census, in particular the extent of undercount of the censuses, in general but more significantly one relative to the other. However, even if the census undercount is low, the census might not identify all the migrants. In general recent migrants are often difficult to include in a census because they have yet to settle. More specifically, immigrants may not be keen to identify themselves as immigrants and either avoid being counted or do not admit to being foreign-born.

Apart from this, place of birth and/or place of residence at previous census, in the case of internal migrants, might be misreported due to boundary changes or ignorance (or even bias) on the part of the respondent.

The third drawback of census data is that it cannot be used to measure emigration from the country of the census. Emigration is particularly difficult to estimate for most countries, but one option is to apply the method for identifying net immigration of the foreigners described above to the censuses of the main countries of destination to which the emigrants move to estimate the change in the numbers of emigrants to those countries. Of course, this is only useful if the censuses of these countries identify the numbers of foreign-born by their countries of birth reasonably accurately.

Generally, statistics on immigrants and particularly emigrants that are collected at border posts provide quite poor estimates of the true numbers, unless the borders of the country are quite impenetrable and there are a few well-controlled ports of entry. Even then there may still be many 'visitors' who end up living in the country.

A final drawback occurs when working with data aggregated over all ages. In these cases one usually has to make use of the crude death rate for the population of the country of the census in order to estimate the number of deaths of the migrant population. However, since the distribution of the migrant population by age can differ from that of the population of the country of the census quite markedly, the estimated number of deaths can be quite inaccurate.

**Extensions of the method**

Some censuses ask additional questions which can be of use in interpreting the patterns of migration, if not improving the estimate of the level of migration. Most common of these is probably a question asking about when the migrant moved. These data allow one to estimate annual rates of migration, however, it is possible that there could be a tendency for respondents to report moves as occurring more recently than is actually the case (Dorrington and Moultrie 2009).

Where a census asks, such as the recent censuses in South Africa, of those who moved since the previous census, where they moved from most recently and when they moved, and

not where they were at the time of the previous census, it is possible to back-project the numbers of migrants by applying annual rates of migration between sub-national regions to estimate the number by place at the time of the previous census (Dorrington and Moultrie 2009). However, in the case of South Africa, at least, it appears that the assumption the most migrants moved only once in the past five years, and thus that the place of residence before the most recent move is the same as the place at the time of the previous census, is quite reasonable (Dorrington and Moultrie 2009).

Where one has data on both the sub-national region of birth and the place at the time of the previous census, one can cross-tabulate the place of residence data by the place of birth and thus be able to classify recent migrants into primary, secondary and return migrants.

## FURTHER READING AND REFERENCES

For general background to the topic of migration, definition of terms and detail on the analysis and interpretation of the data on internal migration the interested reader is referred to the excellent UN manual on the topic, *Manual VI* (UN Population Division 1970). The textbook by Shryock and Siegel (1976) or its modern replacement by Siegel and Swanson (2004) also provides an introduction to the topic of migration and covers, in particular, the estimation of international migration.

Those interested in the estimation of annual migration rates and the back-projection of migration to estimate the numbers by place of residence at the time of the previous census from data on place of residence before the most recent move and year of move are referred to the paper by Dorrington and Moultrie (2009).


Dorrington RE and TA Moultrie. 2009. "Making use of the consistency of patterns to estimate age-specific rates of interprovincial migration in South Africa," Paper presented at Annual conference of the Population Association of America. Detroit, US, 30 April–2 May.

Rogers A and LJ Castro. 1981a. "Age patterns of migration: Cause-specific profiles," in Rogers, A (ed). *Advances in Multiregional Demography (RR-81-006)*. Laxenburg, Austria: International Institute for Applied Systems Analysis, pp. 125–159. http://webarchive.iiasa.ac.at/Admin/PUB/Documents/RR-81-006.pdf

Rogers A and LJ Castro. 1981b. *Model Migration Schedules (RR-81-030).* Laxenburg, Austria: International Institute for Applied Systems Analysis. http://webarchive.iiasa.ac.at/Admin/PUB/Documents/RR-81-030.pdf

Shryock HS and JS Siegel. 1976. *The Methods and Materials of Demography (Condensed Edition).* San Diego: Academic Press.

Siegel JS and D Swanson. 2004. *The Methods and Materials of Demography.* Amsterdam: Elsevier.

Timæus IM. 2004. "Impact of HIV on mortality in Southern Africa: Evidence from demographic surveillance," Paper presented at Seminar of the IUSSP Committee "Emerging Health Threats" HIV, Resurgent Infections and Population Change in Africa. Ougadougou, 12–14 February.

UN Population Division. 1970. *Manual VI: Methods of Measuring Internal Migration.* New York: United Nations, Department of Economic and Social Affairs, ST/SOA/Series A/47. http://www.un.org/esa/population/techcoop/IntMig/manual6/manual6.html

# Chapter 36 The multi-exponential model migration schedule

*Jani Little and Rob Dorrington*

## DESCRIPTION OF METHOD

This chapter describes how to fit a multi-exponential model migration schedule to observed migration data.

Over the last thirty years, these schedules, devised by Rogers and Castro (1981), have been remarkably successful in representing typical age patterns of migration. Essentially the same age patterns of migration have been observed whether national and interregional migrations are considered simultaneously, or migration from a specific region is considered in isolation. The multi-exponential function was designed to reflect the dependency between migration and age, and captures the relationship through an additive sequence of exponential curves, based on 7, 9, 11 or 13 parameters, depending on the complexity of the migration patterns and the ability and robustness of the data to sustain increased parameterization.

When fitted to a schedule of single-year-of-age migration rates, the Rogers-Castro model provides a best-fit, graduated expression of the migration schedule that finds application in smoothing an observed series of migration rates, and which can be used directly to enhance understanding of migration dynamics. The results can also find application in a number of alternative uses, for example, in setting migration schedules to be used in multi-regional population projections. Ideally, the analyst will have estimates of migration by single year and single ages to which the Rogers-Castro model can be fitted. However, if – as is often the case in developing countries where the quality of the underlying data may not permit such finely grained calculations – the data are only available in five-year age groups, then single-year age rates need to be interpolated from the data using one of the methods described in this chapter before attempting to fit a Rogers-Castro model.

## DATA REQUIREMENTS AND ASSUMPTIONS

### Tabulations of data required

- Migration propensities or migration rates by single ages from age 0 to an age above 65 (or if not in single ages, then by five-year age groups).

Ideally the data should be in the form of rates by single ages. Where they are in five-year age groups then single year observations must be interpolated from these five-year estimates before attempting to fit a multi-exponential curve. The choice of the upper age is somewhat arbitrary, but the upper bound of the data used in fitting a model schedule should – at the minimum – be greater than the modal age of retirement.

### Important assumptions

Latest census counts the population by sub-national region and place of birth accurately and identifies who have moved from one region to another since a prior date (e.g. previous census).

## PREPARATORY WORK AND PRELIMINARY INVESTIGATIONS

Before applying this method, you should investigate the quality of the data in at least the following dimensions:

- age structure of the population (by sub-national region as appropriate); and
- relative completeness of the census counts (by sub-national region as appropriate).

## CAVEATS AND WARNINGS

Caution should be exercised in applying the method to net migration data, as the multi-exponential distribution of migration rates by age models gross migration flows (i.e. in- or out-migration) but not necessarily net migration, unless

the flow in one direction significantly dominates the flow in the other at all ages.

## OVERVIEW OF THE MULTI-EXPONENTIAL MODEL MIGRATION SCHEDULE

The multi-exponential function was designed by Rogers and Castro (1981) to reflect the dependency between migration and age. High levels usually found in the first year of life. It drops to a low point during the early teenage years. Then it increases sharply to its highest point during the young adult years. After that, it declines, except for a possible increase and subsequent decrease during the ages of retirement. In some circumstances there may be an upward slope at the oldest ages (Rogers and Castro 1981; Rogers and Watkins 1987).

Over the last thirty years, the schedule (also known as the Rogers-Castro model migration schedule) has proven to be remarkably successful in representing age patterns of migration (Little and Rogers 2007; Raymer and Rogers 2008; Rogers and Castro 1981; 1986; Rogers and Little 1994; Rogers, Little and Raymer 2010; Rogers and Raymer 1999; Rogers and Watkins 1987). These same age patterns of migration have been documented for regions of different sizes and for ethnic and gender sub-populations (Rogers and Castro 1981). They appear whether national interregional migrations are considered simultaneously, or migration from a specific region is considered separately. Directional migration (i.e. from region *i* to region *j*) exhibit the same patterns as well. For example, the Rogers-Castro model migration schedule has been fitted successfully to migration flows between local authorities in England (Bates and Bracken 1982; 1987), Canada's metropolitan and non-metropolitan areas (Liaw and Nagnur 1985), and the regions of Japan, Korea, and Thailand (Kawabe 1990), and South Africa's and Poland's national patterns (Hofmeyr 1988; Potrykowska 1988).

When fitted to a schedule of single-year-of-age migration rates, the Rogers-Castro model provides a best-fit, graduated expression of the migration schedule that can be summarized by 7, 9, 11 or 13 parameters depending on the complexity of the schedule and strength of the data. In addition, the erratic fluctuations, often associated with unreliability in observed age-specific rates, are smoothed.

Rogers-Castro model migration schedules have been used in population projections in Canada (George 1994), and they have been imposed on time periods, regions, and subpopulations (Rogers, Little and Raymer 2010) when migration data were inadequate or unavailable.

The full model schedule has 13 parameters, which is the complete and most complex multi-exponential form of the model. If $M(x)$ is defined as the migration rate for a single year of age $x$, the full model is defined as

$$\begin{aligned} M(x) = {} & a_1 \exp(-\alpha_1 x) \\ & + a_2 \exp\left\{-\alpha_2 (x-\mu_2) - \exp\left[-\lambda_2 (x-\mu_2)\right]\right\} \\ & + a_3 \exp\left\{-\alpha_3 (x-\mu_3) - \exp\left[-\lambda_3 (x-\mu_3)\right]\right\} \\ & + a_4 \exp(\lambda_4 x) \\ & + c \,. \end{aligned}$$

It comprises five additive components. The first component, $a_1 \exp(-\alpha_1 x)$, is a single negative exponential curve representing the migration pattern of the *pre-labour force* ages. The second component, $a_2 \exp\left\{-\alpha_2 (x-\mu_2) - \exp\left[-\lambda_2 (x-\mu_2)\right]\right\}$, is a left-skewed unimodal curve describing the age pattern of migration of people of *working age*. The third component, $a_3 \exp\left\{-\alpha_3 (x-\mu_3) - \exp\left[-\lambda_3 (x-\mu_3)\right]\right\}$, is an almost bell-shaped curve representing the age pattern of migration *post-retirement*, where migration increases sharply following retirement before falling off again. Associated with this component, the fourth component is a single positive exponential curve of the *post-retirement* ages, $a_4 \exp(\lambda_4 x)$, reflecting the (sometimes) observed generalized increase in migration post-retirement. This can be seen, for example, in the migration of the elderly in the US from the North-East to the "sunbelt" states in the South East and South West. The final component is a constant term, $c$, that represents 'background' migration.

Four families of multi-exponential schedules have been identified in past studies (Rogers, Little and Raymer 2010), and only one, exhibiting both a retirement peak and a post-retirement upslope, requires all 13 parameters and all five components. This family is documented in studies of elderly migration (Rogers and Watkins 1987), and is demonstrated in the bottom right panel of Figure 36.1.

The other families are reduced forms of the full model, which means that at least one component is omitted. For example, the most common schedule identified by Rogers, Little and Raymer (2010) requires seven parameters and consists of the first two components and the constant term. This is also called the *standard* schedule, and its shape is set out in the top left panel of Figure 36.1.

A number of schedules have exhibited a standard profile

plus a retirement peak (Rogers and Castro 1981; Rogers and Castro 1986), resulting in the 11-parameter model, including components 1, 2, 3 and 5, shown in the bottom left panel of Figure 36.1. In populations with significant migrant labour, particularly in the developing world, it is possible that the third component is a trough rather than a peak, as migrants return home to retire.

The 9-parameter model is used when the standard pattern is visible for the labour and pre-labour force ages, and there is an upslope to represent migration in the post-retirement years as displayed in the top right panel of Figure 36.1. This was found in several regions of the Netherlands in 1974 by Rogers and Castro (1981).

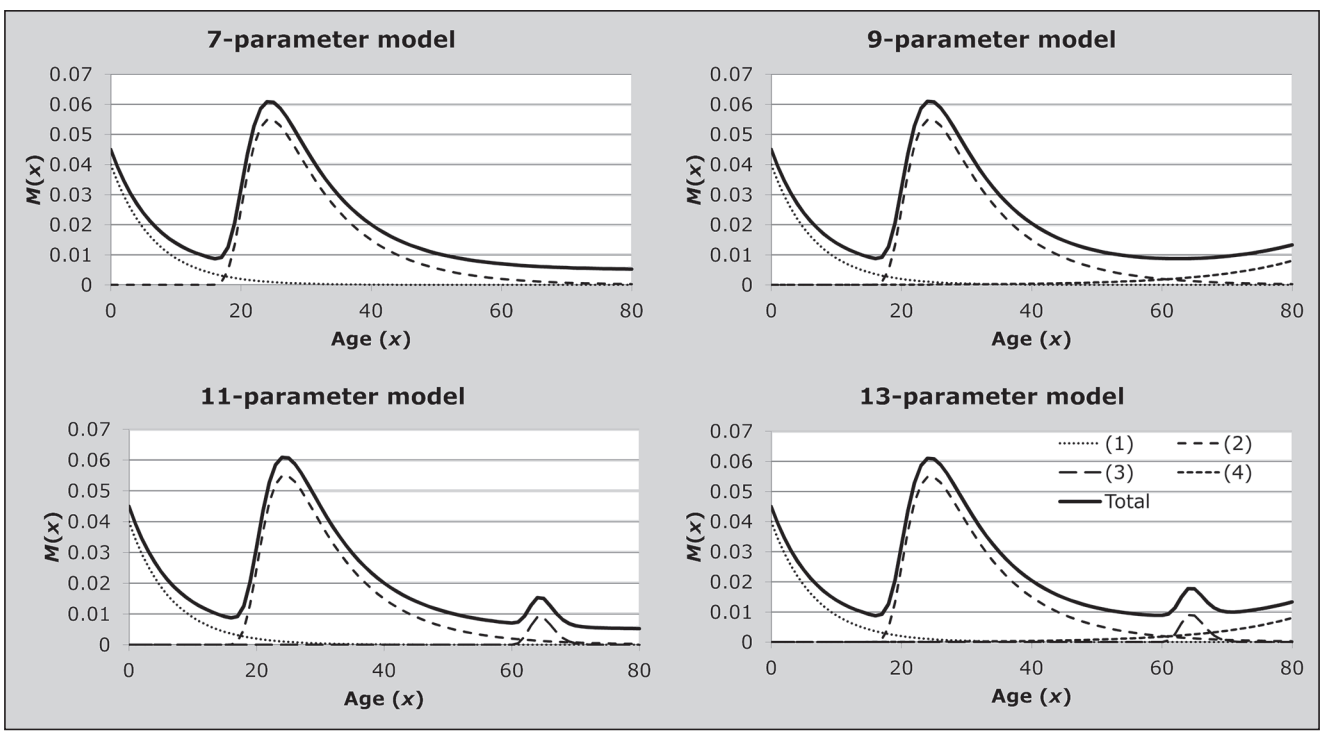
As should be evident from the discussion above, all parameters are interpretable and can be used to characterize the model schedule.

In their original 11-parameter specification of the multi-exponential migration model, Rogers and Castro (1981) illustrated the model using data on male out-migration rates from Stockholm in 1974. Figure 36.2 shows the original data (the jagged lines) and the smoothed 11-parameter schedule fitted to the original data.

Five of the 11 parameters ($\alpha_1$, $\alpha_2$, $\alpha_3$, $\lambda_2$ and $\lambda_3$) give rates of change for different pieces of the model schedule while the level parameters ($a_1$, $a_2$, $a_3$ and $c$) correspond to the heights of the model schedule. $a_1$ gives the peak in the first year of life, $a_2$ is the peak of labour force migration, $a_3$ is the peak of retirement migration, and $c$ gives the background migration rate. $\mu_2$ and $\mu_3$ give the ages at the labour force peak and at the retirement peak, respectively.

Some measures can be used to describe either the observed or the model migration schedule. For example, $x_l$ is the pre-labour force age when migration is at its low point. $x_h$ is the age when labour force migration peaks, and $x_r$ is the age of peak retirement migration. The difference between $x_l$ and $x_h$ is called the 'labour force shift', $X$, and the increase

*FIGURE 36.1 The four main families of the Rogers-Castro model migration schedules showing additive components*




Source: Based on Raymer and Rogers (2008)
Note: The legend indexes, in order, (1) the pre-labour force migration schedule; (2) the working age migration schedule; (3) the post-retirement migratory increase and decrease; and (4) the generalized increase in post-retirement migration.

in migration rate between $x_l$ and $x_h$ is called the 'jump', *B*. *A*, the 'parental shift', is used to describe the average age difference between parent migration and the corresponding migration of children. The gross 'migraproduction' rate (*GMR*) is the sum of all rates over all ages (i.e. the area under the curve), and it is used to gauge the total level of migration out of a region or the total directional migration, i.e., from region *i* to region *j* (Rogers and Castro 1981).

## APPLICATION OF METHOD

The method is applied in the following steps.

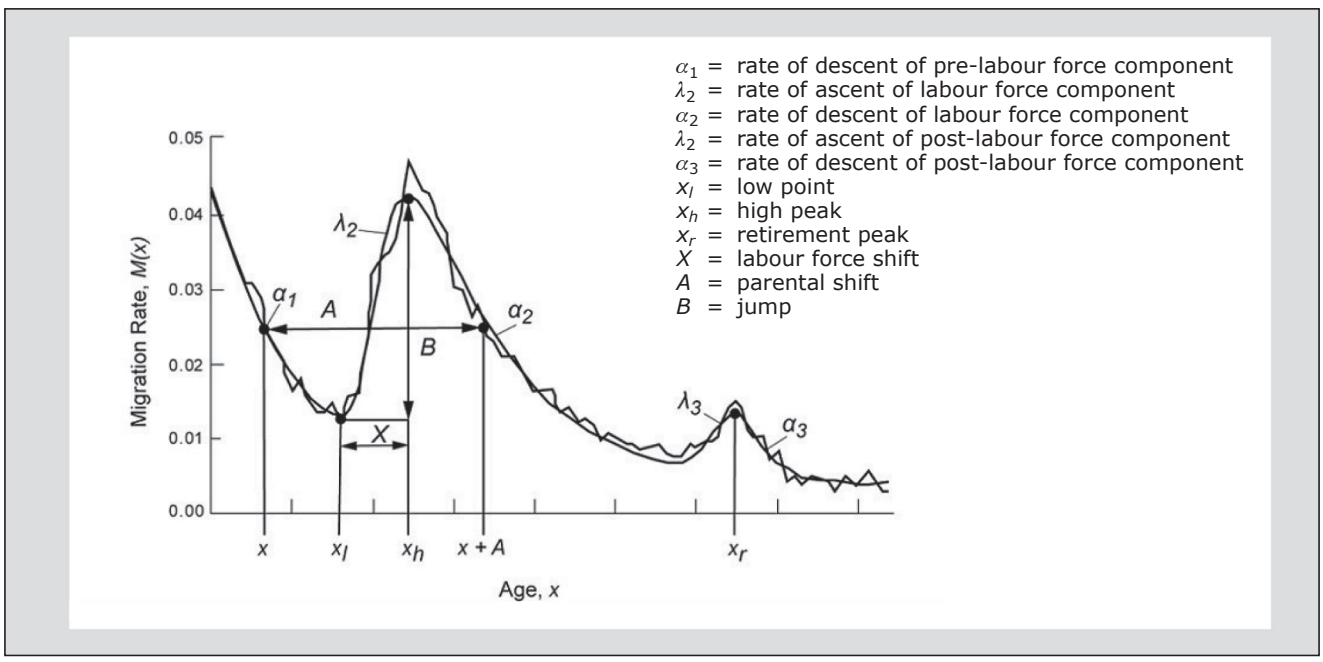
### *Step 1: Prepare a schedule of observed rates*

The initial step in estimating a model schedule is to prepare the data. Decisions about which measure of migration to use depend upon the data sources available (registry, census, or survey) and the purpose of the research. For example, in a comparative study of migration patterns, any of the measures would be appropriate as long as they are constructed similarly across contexts. If, on the other hand, the model schedules are to be used in single-year population projections, the fitted schedule should represent single-age, single-year migration rates. However, where one does not have single-year single-age observations that produce progress relatively smoothly by age, then one must first convert the data one has into single-year single-age estimates. A number of commonly-encountered situations are described below.

#### *A) CENSUS DATA, ANNUAL MIGRATION RATES, FIVE-YEAR MIGRATION INTERVAL*

When the numbers of migrants who survived a five-year migration interval are available from census data which also give the year of most recent move, single-year, single-age migration rates can be derived through a conceptually simple, yet algebraically complex, back-projecting procedure outlined by Dorrington and Moultrie (2009). Their method compensates for the effect of mortality by applying the mortality regime of the general population to the migrants and for the effect of interregional migration by applying the annual rates of migration for the most recent year to estimate

*FIGURE 36.2 The Rogers-Castro 11-parameter model migration schedule fitted to the out-migration rates of males leaving the Stockholm region, 1974*




Source: Rogers and Castro (1981). Permission to reproduce this figure granted by the International Institute for Applied Systems Analysis (IIASA)

the population by region one year prior to the census and using that to estimate the migration rates two years before the census, and using that to estimate the population two years before the census, etc. It requires additional region-of-birth information for those aged 0-4 at time of census, as well as single-age, yearly estimates of regional populations. Schedules derived in this manner can then be fitted and smoothed with a Rogers-Castro model schedule, and used in single-year population projections.

#### *B) INTERPOLATING ONE-YEAR FROM FIVE-YEAR AGE PROPENSITIES*

Regardless of the migration time interval, whether using census data or population register data, five-year age groupings generally give more reliable estimates of migration propensities than one-year age categories (Rogers, Little and Raymer 2010). In addition, counts of migrants in one-year age categories are typically only available from sample data, since national population bureaus tend to publish counts of interregional migrants in five-year age categories.

To apply the multi-exponential model when the initial migration proportions are in five-year age categories requires some method of converting the five-year rates to one-year rates. Cubic-spline interpolation (McNeil, Trussell and Turner 1977) is one such method that produces a smooth schedule for all integer values of ages. Rogers and Castro (1981) used data from Sweden, which was available in one-year and five-year age rates, to test the accuracy of the cubic-spline method, and found generally satisfactory results.

To arrive at smooth one-year age migration profiles, the initial migration proportions for the five-year age categories are assigned values close to the middle age within the five-year interval, i.e., ages 2, 7, 12, 15, … 72, 77 (or 2.5, 7.5, 12.5, …, etc., if estimating rates rather than probabilities). From this set of points, a continuous age profile of state outmigration propensities is generated with cubic-spline interpolation, which constructs third-order polynomials that pass through the set of pre-defined control points (called nodes). Commercial or freeware add-ins for Microsoft Excel, such as *XlXtrFun*, can also be used to implement cubic spline interpolation.

An alternative approach is to adapt Beers' 6-parameter interpolation procedure (Beers 1945) to interpolate rates between the rates for the youngest and oldest age groups, which also extrapolates the rates to ages 0 and 1 (or 0.5 and 1.5 if working with rates). The extrapolation to the youngest ages is achieved by assuming that the difference between propensities for age 1 and 2 is the same as that between ages 2 and 3, and that between ages 0 and 1 is the same as that between ages 3 and 4.

Thus, to apply either approach one needs a set of migration rates in five-year age intervals from 0-4, to at least 65-69.

***Step 2: Decide on the form of the multi-exponential model***

Once the observed schedule is prepared, a decision must be made about the form of the multi-exponential model to be adopted. The overview of the multi-exponential model migration schedule presented above described the characteristics of the 7-, 9-, 11-, and 13-parameter models. This decision should be informed by a visual inspection of the schedule, keeping in mind that the model is assumed to represent the true form of the population migration schedule. Sometimes, even after plotting the schedule, it is not apparent how best to model the retirement years and the oldest ages. For example, it may appear that either a standard 7-parameter model or a 9-parameter model (increasing migration in the oldest ages) would be appropriate. In this situation, the decision in favour of the 9-parameter model could be based on a theoretical expectation for increasing migration in the later years. On the other hand, the 9-parameter model form might be rejected, based on the goodness-of-fit measures, as being insufficiently parsimonious if it produces no better fit than the 7-parameter model. In deciding which form of the model to use, it is recommended that the goodness-of-fit of the simpler model be compared with the more complex model, (e.g. comparing the fit of a 7-parameter model versus that of an 11-parameter model). As a general rule, and always bearing in mind the likely robustness of the underlying data, substantial improvement in fit is required to justify a more complex specification.

For most developing countries, particularly where 'retirement' isn't concentrated between the ages of 60 and 65 and there is age exaggeration at the older ages, the data are probably not strong enough to fit anything more than the 7-parameter version of the model.

***Step 3: Fit the model using Solver***

Given the number of parameters (between 7 and 13) in the multi-exponential model migration schedule, determining a best fit *ab initio* using trial-and-error is not recommended. Instead, analytical algorithms have to be employed. The one described below uses an algorithm that is provided in

Microsoft Excel, *Solver. Solver* may not be routinely loaded by standard installations of Microsoft Excel. To enable its use, proceed by selecting "File → Options → Add-ins → Manage Excel Add-ins → Go …" and then ensuring that the "Solver Add-in" is ticked.

The specifications of the *Solver* function, and the conditions and constraints that should be adhered to, have been set up in the workbook associated with the methods presented in this chapter. To run the routine on a given worksheet, select "Data → Solver → Solve".

The model is fitted in the associated workbook and is set up to allow the user to set the "objective" to be minimized to be either the sum of squared differences between the observed rates and the fitted rates, or the chi-squared statistic.

The default *Solver* is set up to fit using all parameters. If one wants to fit a curve using only some of the parameters then one must specify only these parameters in the "By Changing Variable Cells" window, and set the other parameters to appropriate constant values (which may, or may not, be zero depending on the requirements of the fitting procedure). An instance where such constrained optimization may be required is mentioned below.

The sum of squared differences is calculated as follows:

$$\frac{1}{n}\sum_{i=1}^{n}\left(O_i - F_i\right)^2$$

where *Oi* represents the observed rate at age *i*, *Fi* represents the fitted value at age *i* and *n* the number of age groups.

The chi-squared statistic is calculated as follows:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\left(O_i - F_i\right)^2}{F_i}.$$

The chi-squared statistic is more sensitive to misfitting to age ranges where rates are lower (resulting in a proportionately larger error) and thus is a better metric to assess goodness-of-fit when trying to fit the 'retirement hump' (the third component).

#### *CHOOSING INITIAL ESTIMATES FOR THE FITTING PROCEDURE*

The choice of initial parameter values is the principal difficulty in non-linear parameter estimation. Ideally, given a set of starting values, the algorithm proceeds through an iterative process, producing a revised set of "optimum" values. However, the optimum may be merely a local optimum, and not the global optimum. A better guess of the initial parameter values may produce an improved goodness-of-fit and produce a different set of final values. A poorer choice of initial parameters may prevent convergence to even a local optimum.

Bearing this in mind, the most effective method of ensuring that the results from a fitting procedure are indeed globally "optimal" is to choose parameter values previously reported for a "similar" curve. To this end one might start with the values already in the workbook which were used to fit the curves in the examples below.

Convergence may be more difficult to achieve with the 11- and 13-parameter models. Where such heavily parameterized models are justified, one approach that can be adopted is to first fit a standard 7-parameter model to the data (thereby securing the fit at the peak of the migration schedule, and at ages up to mid-adulthood). Then, one could proceed by fixing those 7 parameters to their estimates that resulted from the initial step (i.e. treat those parameters as constant from there on), and then estimate the remaining parameters. Another effective procedure is to carry out a linear estimation method first, which does not rely on an iterative algorithm. That method was first described in Rogers and Castro (1981) and later included as one of the several alternatives set out in Rogers, Castro and Lea (2005).

Another challenge in finding the optimum solution lies in choosing an appropriate stopping criterion for the iterative algorithm. As the iteration process converges on a solution, the chi-square statistic, which measures the differences between the observed and the estimated values, decreases. An indication that an acceptable solution has been found is when the chi-square value decreases by only a negligible amount from one iteration to the next. The level of this small difference is called the "tolerance" and is set by the user. The temptation is to set it to be a very small value, i.e. very close to zero, so that a true minimum chi-square value is achieved. However, the risk in this approach is that such a low tolerance may not be achievable, even when a solution has been found. Press, Flannery, Teukolsky *et al.* (1986) suggest a tolerance equal to 0.001 is a reasonable setting. If the estimation software fails to converge, the convergence criteria could be made less stringent, i.e. increase the tolerance, or try new initial estimates.

One trial-and-error method of choosing initial estimates makes use of the graphs in the accompanying Excel workbook (see website). By substituting your schedule of observed data in one of the sheets, initial "guesses" of each parameter can be chosen and placed in the cells where the final estimates

of each parameter are located. Then, by visual examination of the fit, and identification of the parameter values that are most out of line, try new initial values for those parameters and then re-evaluate the fit visually. Continue this way until the fitted schedule is reasonably close to the observed schedule. At this point, you will know you have reasonably good initial estimates and may proceed to the nonlinear least squares estimation procedure.

***Step 4: Evaluate the model fit***

We evaluate the model fit by calculating the mean absolute percent error (*MAPE*) statistic:

$$MAPE = 100\frac{1}{n}\sum_{i=1}^{n}\left[\frac{|F_i - O_i|}{O_i}\right].$$

The *MAPE* is prone to overstate inaccuracy, particularly when the observed schedule has many values that are very close to zero (Morrison, Bryan and Swanson 2004).

In addition to *MAPE*, we also calculate $R^2$, the square of the correlation between the $O_i$ and the $F_i$ values. A heuristic that is often employed is that a reasonable fit is achieved with a *MAPE* of 15 per cent or less together with an $R^2$ well above 90 per cent.

In addition, since the method assumes the estimated Rogers-Castro model schedule represents the true form of the migration schedule, the estimated model schedule should appear to represent the underlying pattern of the observed data.

***Step 5: Interpret the results of the fit***

If the goal is to describe the pattern of migration and a multi-exponential model has been successfully fitted to the data, any of the summary measures (e.g. *GMR*, *X*, *B*, and *A*) as well as the parameter estimates can be used to describe the schedule. The summary measures and the parameter interpretations are given in the Overview presented earlier in this chapter.

## WORKED EXAMPLES

In the examples below, multi-exponential model migration schedules are applied to a variety of data, of varying quality and complexity and from a number of different sources. All worked examples are provided in the associated workbook on the *Tools for Demographic Estimation* website.

Because iterative methods are required to fit a model life table to data on conditional survivorship in adulthood, detailed worked examples are not provided in the text. The reader is directed to the description provided in the previous section on how to use *Solver* in Microsoft Excel to determine optimal fits. The workbook is set up to use *Solver* to derive the results presented.

**Census data, one-year migration interval**

An example of a schedule based on one-year age migration propensities measured over a one-year migration interval from census data is shown in Figure 36.3. The data are derived from the 2005 American Community Survey (ACS), a national survey conducted annually by the US Census Bureau. Even for California, a highly populated state, the one-year age propensities over a one-year interval are quite unstable. The *MAPE* is 17 per cent and the $R^2$ is 0.92.

Caution must be exercised when using one-year age propensities over one-year migration intervals. For each single age, the numbers at risk of migrating, as well as the numbers of migrants, may be small, resulting in propensities that are erratic and unstable. A better option may be to derive five-year age propensities, which have proven to be more reliable than one-year age propensities (Rogers, Little and Raymer 2010). These can be interpolated to yield one-year age propensities using cubic splines or Beers' formula as discussed in the section describing the application of the method.

**Census data, five-year migration interval**

Figure 36.4 shows an example using census data for the state of New Hampshire. The US Census Bureau's 1 per cent Public Use Microdata Sample (PUMS) is a relatively small sample taken from the census and New Hampshire is one of the least populated states. The one-year age propensities appear to be quite unstable with dramatic fluctuations, while the model schedule provides a smooth estimate of the true schedule form. The *MAPE* is 52 per cent and the $R^2$ is 0.68.

Figure 36.5 shows the cubic-spline interpolation method applied to the five-year age migration propensities for New Hampshire, derived from the 2000 Census 1 per cent public use microsample data. The schedule interpolated from the five-year age rates is much smoother and provides more reliable estimates than the observed one-year age rates displayed in Figure 36.4, and thus is a better set of estimates against which to compare the fitted multi-exponential curve. The *MAPE* was reduced from 52 per cent for the one-year age propensities to 15 per cent for the rates interpolated

from the five-year age proportions, and the $R^2$ increased from 0.68 to 0.94.

There are several reasons why the levels of the New Hampshire schedule, in Figure 36.5, are substantially higher than the California schedule, Figure 36.4. The California example gives migration over a one-year migration interval and the New Hampshire schedule is over a five-year interval. In addition, New Hampshire is a much smaller areal region than California and the expectation is that the force of migration will be more powerful in a geographically smaller region.

## DIAGNOSTICS, ANALYSIS AND INTERPRETATION

### Checks and validation

It is important to check visually if the age-specific migration rates have a 'shape' that is compatible with the Rogers-Castro models. If this is not the case then it is unlikely that these models will provide a satisfactory fit. Likewise, it is worthwhile checking whether there are any extreme values, particularly at older ages which might distort the choice of parameters or even the choice of the number of parameters to be fitted. If the observed estimates are particularly noisy, it would be better to group the data into five-year age intervals and then estimate a smoothed distribution using either the Beers 6-parameter interpolation provided or Spline curve fitting.

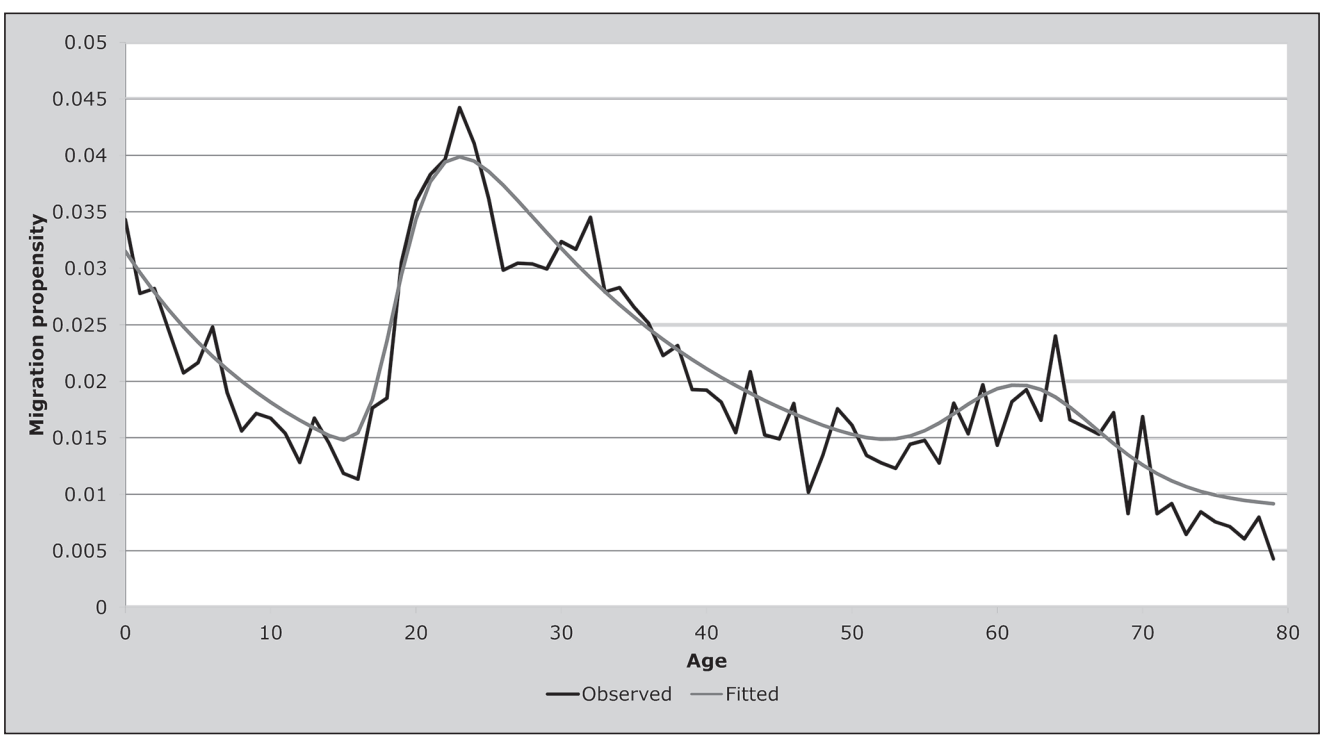
## CONSIDERATIONS IN APPLYING THE METHOD

The formulation of the multi-exponential model was presented in the Overview presented earlier in this chapter and is not repeated here. In this section, we discuss in greater detail aspects that should be considered carefully before applying the method in practice.

### Data preparation

The multi-exponential model is applied to schedules of one-year age migration rates beginning at age 0 and, typically, continuing to age 65 or higher to capture the full pattern of elderly migration. The schedules of age-specific migration might measure directional migration (i.e. from region $i$ to region $j$) or total out-migration (i.e. from region $i$ to all other regions), or all inter-regional migration with no specific origin or destination. Usually, migration data are

*FIGURE 36.3 Out-migration over the 1-year interval, 2004–2005, with fitted 11-parameter model schedule, California, 2005 ACS*



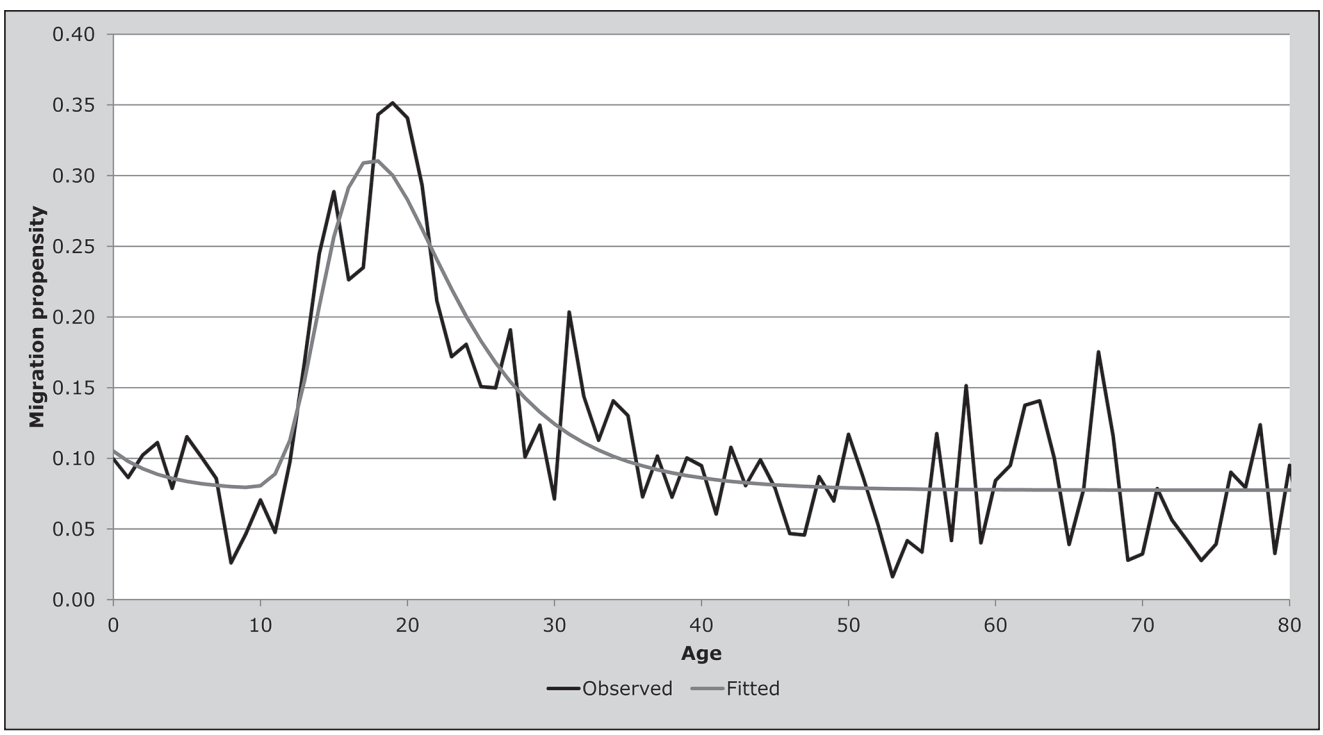
obtained from national censuses (or, in developed countries, population registers). The multi-exponential model can be applied to a variety of measures of single-age migration propensities derived from either of these sources.

When obtained from national registration systems, the migration rate, for persons aged $x$ at the beginning of the interval, is the ratio of the number of migrations during a given time interval divided by the average number of person-years exposed to the risk of moving. Persons can contribute more than one migration during the interval. These are occurrence-exposure rates, although migrations by non-survivors may not be included in the numerator (Rogers and Castro 1981).

The observed migration schedule in Figure 36.2 was derived from Sweden's national registry for male migration out of Stockholm over a one-year interval. In contrast, Figure 36.6 shows the observed and estimated model schedule for all male inter-communal migration in Sweden over a five-year interval. As expected, the levels are much higher in Figure 36.6 due to more migration activity when all regions are combined as compared to the Stockholm region alone. Similarly, more migrations are expected over a five-year interval than over a one-year interval. Rees (1977) found migration rates over a five-year interval tend to be less than five times (between three and five times) those over a one-year interval. The observed schedule is also smoother and more similar to the model schedule in Figure 36.6, indicating single-age migration rates are more reliable when based on a longer interval.

Censuses, on the other hand, count surviving migrants (not migrations). Migrants are persons who reported living in one region, at the beginning of the time interval, and resided in a different region at the time of the census. A person registering multiple migrations in a national register may be a non-migrant in the census if he returned to his initial location during the time interval. In general, counts of migrants from censuses understate the number of migrations, especially for longer time intervals when there are bound to be larger numbers of return movers and non-survivors. For these reasons, a migration schedule derived

*FIGURE 36.4 Out-migration over the 5-year interval, 1995–2000, and fitted with a 7-parameter model schedule, New Hampshire, 2000 Census 1 per cent PUMS*



from population register data is not directly comparable to one based on census data (Rogers and Castro 1986).

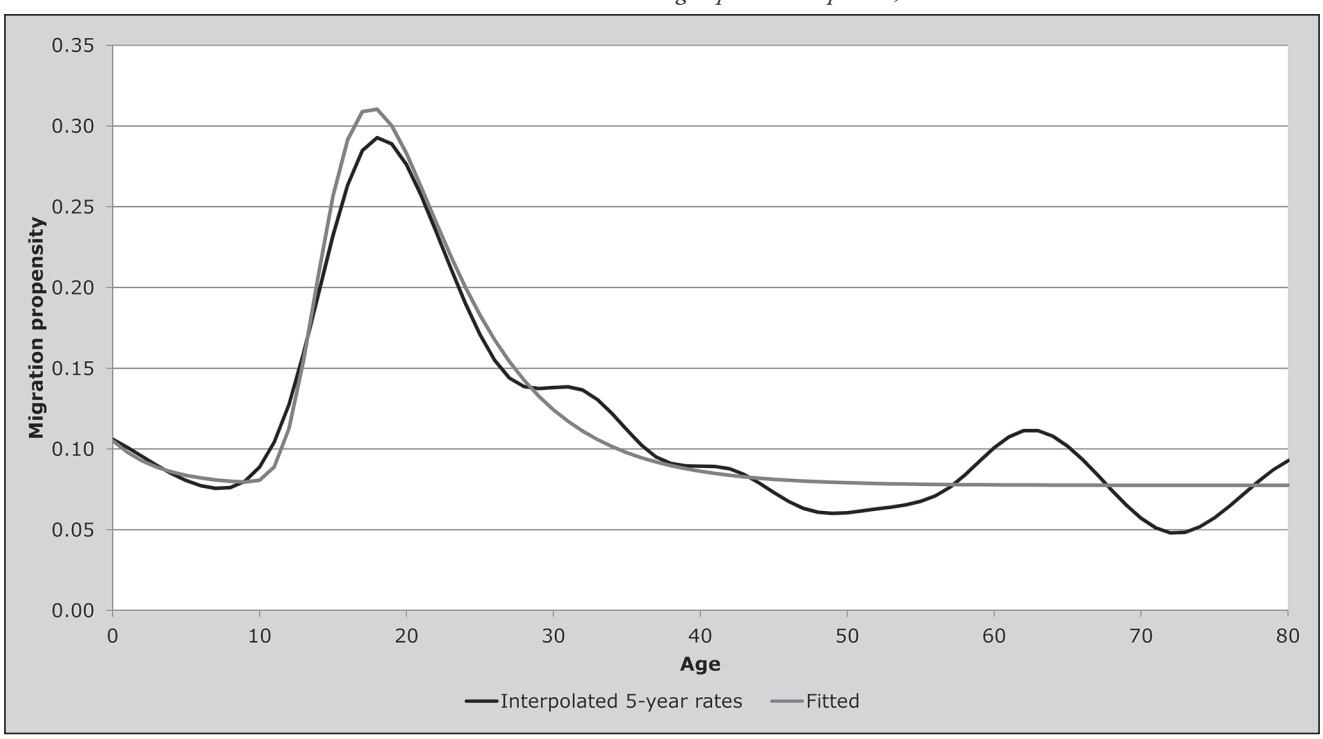
Censuses typically record the location of a person's current residence and ask where the person was living either one year ago or five years ago. Given this information and the person's age at the time of census, the numbers of surviving migrants, and the numbers of survivors who were at risk of migrating are counted. The ratio of the number of surviving migrants to the number of survivors at risk for migrating is sometimes called a 'conditional survivorship proportion' because migrants and persons counted as being at risk for migrating must have survived the migration time interval to be counted by the census (Rogers, Little and Raymer 2010). Since these are not occurrence-exposure rates they will be called migration propensity here.

***Census data, one-year migration interval***

To derive single-age migration propensities when the census question asks where a person was living one year ago, all persons are "back-cast" to the region where they lived one year earlier when they were one year younger, which gives the number of persons at risk of migrating from that region. For example, a person aged 1 last birthday in a census conducted in 2010 would have been aged 0 last birthday in 2009. If the 2010 age values ranged from 1 to 85, they would range from 0 to 84 in 2009. (Note, only persons aged 1 and older would have reported place of residence 1 year ago.) Back-casting yields the number of people who survived to be counted by the census in 2010 and who were at risk for migrating from region *i*, in 2009. The number of migrants would be the count of persons who reported living in region *i* in 2009, but were counted as residing in a different region in 2010. For each 1-year age group, the ratio of the number of migrants to the number at risk for migrating gives the age-specific out-migration propensity for the 1-year interval. When the numerator contains directional migrants, i.e. from region *i* to region *j*, the ratio gives the age-specific propensity to migrate from region *i* to region *j*.

Caution must be exercised when using one-year age propensities over one-year migration intervals. For each

***FIGURE 36.5*** *Out-migration over the 5-year interval, 1995–2000, and fitted with a 7-parameter model schedule, New Hampshire, 2000 Census 1 per cent PUMS. (Migration rates in 5-year age groups and interpolated)*



single age, the numbers at risk of migrating, as well as the numbers of migrants, may be small, resulting in propensities that are erratic and unstable. A better option may be to derive five-year age propensities, which have proven to be more reliable than one-year age propensities (Rogers, Little and Raymer 2010). These can be interpolated to yield one-year age propensities.

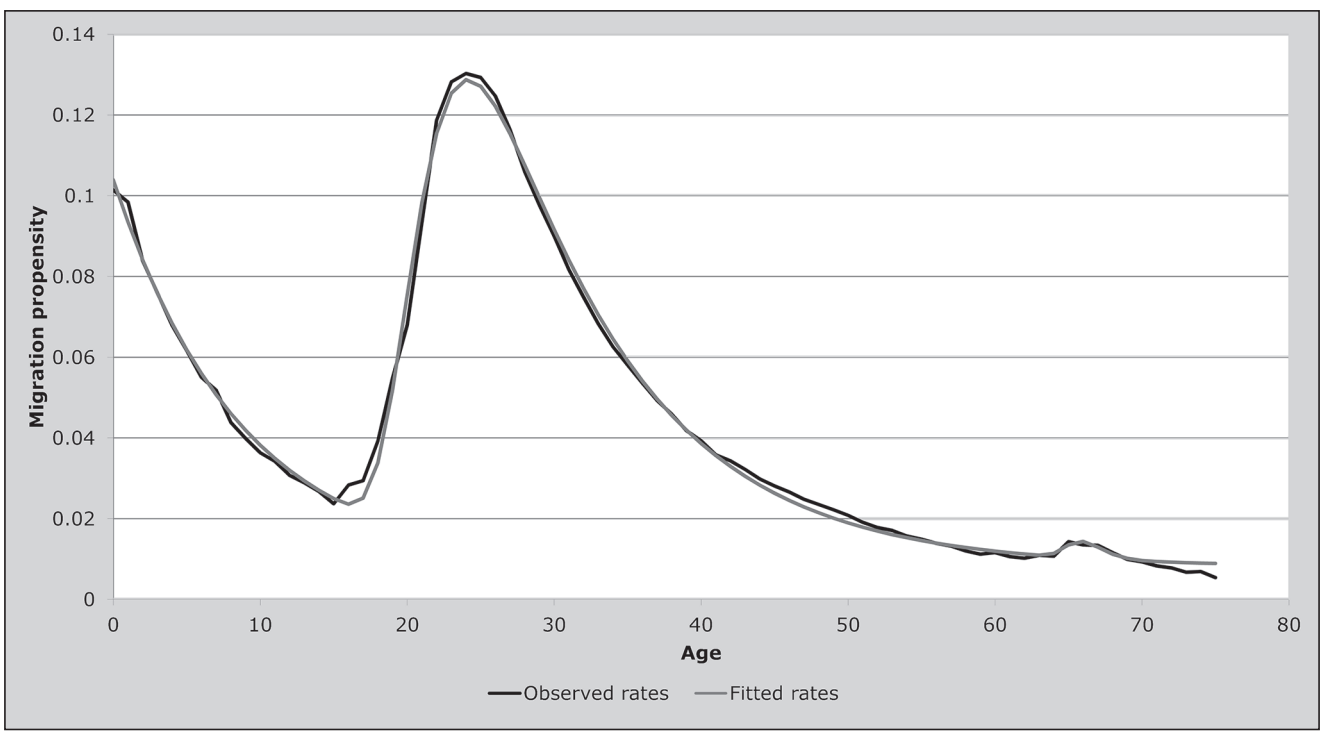
***Census data, five-year migration interval***

When the census question asks where a person was living five years ago, it is possible to derive one-year age propensities for migrating over a five-year interval as long as single ages are reported. It is done by back-casting all persons to the region where they lived five years earlier when they were five years younger. Persons aged 5 last birthday in a census conducted in 2000, for example, would have been aged 0 last birthday in 1995. If the age values ranged from 5 to 85 in 2000, they would range from 0 to 80 in 1995. The number of migrants is simply the count of persons who reported living in region *i* in 1995, but were counted as residing in a different region in 2000. For each one-year age group, the ratio of the number of migrants to the number at risk for migrating gives the age-specific out-migration propensity over the five-year interval.

***Census data, annual migration rates, five-year migration interval***

When the numbers of migrants who survived a five-year migration interval are available from census data, single-year, single-age migration rates can be derived through a back-projecting procedure outlined by Dorrington and Moultrie (2009). Their method compensates for the effect of mortality by applying the mortality regime of the general population to the migrants and for the effect of onward migration by applying the annual rates of migration for the most recent year to estimate the population by region one year prior to the census and using that to estimate the migration rates two years before the census, and using that to estimate the population two years before the census, etc. It requires additional region-of-birth information for those aged 0–4 at time of census, as well as single-age, yearly estimates of

*FIGURE 36.6 Swedish male inter-communal migration over the 5-year interval, 1968–1973 with fitted 11-parameter model schedule*



regional populations. Schedules derived in this manner can then be fitted and smoothed with a Rogers-Castro model schedule, and used in single-year population projections.

**Limitations**

Unless one has accurate and well-behaved data the multi-exponential model will not produce a very close fit and thus can be over-parameterized – i.e. many different sets of parameters can produce virtually equally good fits to the observed values. In such a situation it might help to fix one or two parameter values and fit the rest, and parsimony with the number of parameters is recommended.

**Extensions**

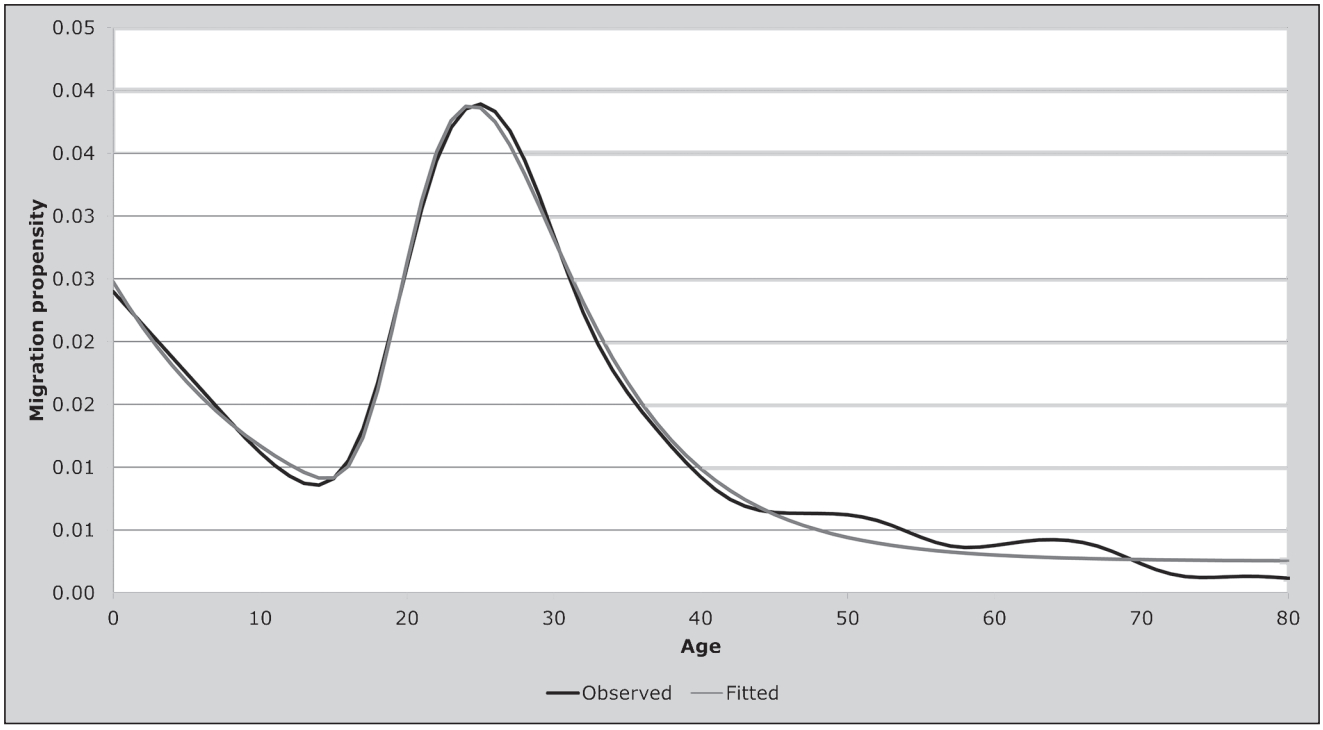
Application of the multi-exponential model is not limited to schedules of migration rates or propensities. Several studies have established that age distributions of migrants (and migrations if using registration data) often have a multi-exponential form and can be accurately represented by a Rogers-Castro model schedule (Little and Rogers 2007; Rogers, Little and Raymer 2010).

The single-age numbers of migrants/migrations can be derived using any of the data sources and methods described above, because these are simply the numerators in the migration propensity and rate calculations. The observed data fitted by the model schedules are the single-age proportions of the total migrants/migrations. Note, if the numbers of migrants are reported in five-year age categories, some form of interpolation would be necessary. If cubic spline interpolation is used, the numbers associated with each node should be the migrants/migrations for each five-year age grouping divided by five.

For example, the observed age composition of Swedish migrations as a proportion is illustrated in Figure 36.7. From this it appears to be very smooth and reliable except in the oldest ages. A 7-parameter model schedule fits pretty closely, with an $R^2$ of 99 per cent and *MAPE* of 29 per cent. However, this is an example of the how the *MAPE* can exaggerate the model's lack of fit, as it becomes inflated when there is a sequence of small observed deviations.

Two alternative software options for fitting to the Excel workbook for fitting the multi-exponential curve are 1) *Data*

***FIGURE 36.7*** *Swedish age composition of inter-communal migration, 1968-1973 and fitted with a 7-parameter model schedule*



*Master 2003*, a free curve-fitting program, which applies the Levenberg–Marquardt algorithm; and 2) *R* (R Development Core Team 2012) which is also free, but is a software environment for all-purpose statistical computing and graphics and as such requires a significant time investment before it can be used with confidence. The Appendix to this chapter on the *Tools for Demographic* Estimation website gives very basic commands for defining *R*-functions that produce estimates for the 7-parameter and the 11-parameter models using the Gauss-Newton algorithm.

## REFERENCES


Bates J and I Bracken. 1982. "Estimation of migration profiles in England and Wales", *Environment and Planning A* **14**(7):889–900. doi: 10.1068/a140889

Bates J and I Bracken. 1987. "Migration age profiles for local-authority areas in England, 1971–1981", *Environment and Planning A* **19**(4):521–535. doi: 10.1068/a190521

Beers HS. 1945. "Six-term formulas for routine actuarial interpolation", *The Record of the American Institute of Actuaries* **33**(2):245–260.

Dorrington RE and TA Moultrie. 2009. "Making use of the consistency of patterns to estimate age-specific rates of interprovincial migration in South Africa," Paper presented at Annual Meeting of the Population Association of America. Detroit, MI, 29 April–2 May 2009.

George MV. 1994. *Population projections for Canada, provinces and territories, 1993–2016.* Ottawa: Statistics Canada, Demography Division, Population Projections Section.

Hofmeyr BE. 1988. "Application of a mathematical model to South African migration data, 1975–1980", *Southern African Journal of Demography* **2**(1):24–28.

Kawabe H. 1990. *Migration rates by age group and migration patterns: Application of Rogers' migration schedule model to Japan, The Republic of Korea, and Thailand*. Tokyo: Institute of Developing Economies.

Liaw K-L and DN Nagnur. 1985. "Characterization of metropolitan and nonmetropolitan outmigration schedules of the Canadian population system, 1971–1976", *Canadian Studies in Population* 12(1):81–102.

Little JS and A Rogers. 2007. "What can the age composition of a population tell us about the age composition of its out-migrants?", *Population, Space and Place* **13**(1):23–19. doi: 10.1002/psp.440

McNeil DR, TJ Trussell and JC Turner. 1977. "Spline interpolation of demographic data", *Demography* **14**(2):245–252. doi: 10.2307/2060581

Morrison PA, TM Bryan and DA Swanson. 2004. "Internal migration and short-distance mobility," in Siegel, JS and DA Swanson (eds). *The Methods and Materials of Demography*. San Diego: Elsevier pp. 493–521.

Potrykowska A. 1988. "Age patterns and model migration schedules in Poland", *Geographia Polonica* **54**:63–80.

Press WH, BP Flannery, SA Teukolsky and WT Vetterling. 1986. *Numerical Recipes: The Art of Scientific Computing.* Cambridge: Cambridge University Press.

R Development Core Team. 2012. *R: A language and environment for statistical computing: Reference Index*. Vienna, Austria: R Foundation for Statistical Computing. http://www.mendeley.com/research/r-language-environment-statistical-computing-13/

Raymer J and A Rogers. 2008. "Applying model migration schedules to represent age-specific migration flows," in Raymer, J and F Willekens (eds). *International Migration in Europe: Data, Models and Estimates*. Chichester: Wiley, pp. 175–192.

Rees PH. 1977. "The measurement of migration, from census data and other sources", *Environment and Planning A* **9**(3):247–272. doi: 10.1068/a090247

Rogers A and LJ Castro. 1981. *Model Migration Schedules.* Laxenburg, Austria: International Institute for Applied Systems Analysis. http://webarchive.iiasa.ac.at/Admin/PUB/Documents/RR-81-030.pdf

Rogers A and LJ Castro. 1986. "Migration," in Rogers, A and F Willekens (eds). *Migration and Settlement: A Multiregional Comparative Study*. Dordrecht: D. Reidel, pp. 157–208.

Rogers A, LJ Castro and M Lea. 2005. "Model migration schedules: Three alternative linear parameter estimation methods", *Mathematical Population Studies* **12**(1):17–38. doi: 10.1080/08898480590902145

Rogers A and JS Little. 1994. "Parameterizing age patterns of demographic rates with the multiexponential model schedule", *Mathematical Population Studies* **4**(3):175–195. doi: 10.1080/08898489409525372

Rogers A, JS Little and J Raymer. 2010. *The Indirect Estimation of Migration: Methods for Dealing with Irregular, Inadequate, and Missing Data.* Dordrecht: Springer.

Rogers A and J Raymer. 1999. "Estimating the regional migration patterns of the foreign-born population in the United States: 1950–1990", *Mathematical Population Studies* **7**(3):181–216. doi: 10.1080/08898489909525457

Rogers A and J Watkins. 1987. "General versus elderly interstate migration and population redistribution in the United States", *Research on Aging* **9**(4):483–529. doi: 10.1177/0164027587094002

# Chapter 37 Log-linear models of migration flows

*Jani Little and James Raymer*

## INTRODUCTION TO MODEL APPLICATIONS

The log-linear modelling framework provides several valuable techniques for studying and estimating migration flows within a network of regions. To date, these methods have been applied most often to internal migration systems where regions are defined as sub-national administrative units. However, they need not be restricted to domestic migration and may be applied to international systems of migration as well (Raymer 2007).

A migration flow is defined as the number of migrations from one region to another over the course of a specified time frame. There are several different ways to count migrations and each one could yield a different result. For example, Rees and Willekens (1986) make the distinction between registration systems that count the number of inter-regional residential moves over a reference period and censuses that count persons who reside in a place at the time of the census that is different from the place of residence at the beginning of the reference period.

Regardless of the method used to count migration flows, it is conventional to present them in contingency tables. These are square tables that report the flow counts between origin and destination regions. The flows in the migration table can be perfectly reproduced by the multiplicative component model, which is a saturated (i.e., where there are as many estimated parameters as there are data points) log-linear model. It has been used by Willekens (1983), Rogers, Willekens, Little *et al.* (2002) and Rogers, Little and Raymer (2010) to represent the matrix of flows between regions, and by Raymer and Rogers (2007), Raymer, Bonaguidi and Valentini (2006) and Rogers, Little and Raymer (2010) to capture the structure of inter-regional flows within age categories. The multiplicative components are interpretable and conveniently used to define the structure of migration between the regions of interest (Rogers, Willekens, Little *et al.* 2002). If calculated for more than one set of inter-regional flows, defined for different time periods, for example, or for different age, sex or race categories, multiplicative components are useful for comparing migration regimes across these populations.

Log-linear methods may be used to justify simplified representations of migration structure that are more parsimonious than the saturated model. The appropriateness of a reduced model is determined by fitting the predicted flows to the observed flows and by using statistical methods to evaluate the goodness of fit. If the reduced form has merit, i.e., fits the data well, the model may be used to estimate indirectly the flows. The independence model, for example, assumes inter-regional flows are distributed according to the pattern that could have been predicted based on the marginal distributions of flows across origin and destination regions. If the independence model is confirmed, inter-regional flows are predictable and can be estimated indirectly, but accurately, if the total sending and receiving flows of each region are given.

Sometimes the structure of migration is hypothesized to be invariant with respect to factors such as time, age, sex, and race. These hypotheses can be represented and tested with log-linear models. Allowing for changes in the level of migration, studies have documented remarkable stability in migration structures, in particular the rates of migration by age, over time (Mueser 1989; Nair 1985; Snickars and Weibull 1977). Other studies have shown consistency in the age patterns of inter-regional migration over time (Raymer and Rogers 2007). Moreover, the migration structure of the youngest ages, which can be inferred from birthplace-specific population stocks, has, in certain contexts, proven to be a "proxy" for the level of migration and allowed the estimation of migration of the older age groups (Raymer and Rogers 2007; Rogers, Little and Raymer 2010)

These studies have set the stage for establishing the method of offsets as a successful tool for indirectly estimating

migration flows. It is a special application of log-linear modelling that forces a known migration structure on to a system that may have missing or unreliable inter-regional flow data. Using this method, the known migration structure of one time period can be borrowed from another period. In addition, when flows are disaggregated by age, the structure of age-specific inter-regional flows of one time period can be applied to another period. Furthermore Raymer and Rogers (2007) showed that the level of infant lifetime migration can be applied, using the method of offsets, to estimate indirectly the migration flows of the older ages.

Applications of log-linear models, and the related assumptions, are detailed in the sections that follow, beginning with the two-variable case, i.e., origin and destination. In this section, the log-linear model is defined in the context of two-dimensional flow tables, and multiplicative forms as well as additive forms of the saturated model are derived and interpreted. The log-linear model of independence and the "migrants only" quasi-independence model are set out, including illustrations and a brief description of the methods for evaluating goodness-of-fit.

The section concludes with an illustration of the method of offsets for indirectly estimating the inter-regional flow data of one period based on the migration flow patterns of another. When flow data are available for two periods, the period-invariance assumption can be tested with a log-linear model and the method of offsets. Models that disaggregate the origin and destination of flows into age categories are considered. This is followed by an illustration of how the multiplicative model with age can be applied, using the method of offsets, to estimate indirectly the age-specific inter-regional flows for another period.

## APPLICATIONS OF THE TWO-VARIABLE MODEL

To illustrate the two-variable log-linear model, consider the 1973 and 1976 migrations in the Netherlands between types of municipalities categorized into six different groups based on degree of urbanization. These were published by Willekens (1983) and are presented in Table 37.1. In this context, there are two variables, region of origin ($O$) and region of destination ($D$). Neither is identified as the dependent variable. The outcome variable may be either the inter-regional migration flow, denoted $n_{ij}$, in the multiplicative form of the model, or the natural logarithm of the flow, denoted $\ln(n_{ij})$, in the additive form of the model.

Decompositions of the saturated model, each one perfectly regenerating the observed data, are described in the subsections presenting the multiplicative component model and the additive linear model, and three indirect estimation techniques are illustrated in the three subsections describing the independence model, the quasi-independence model and the method of offsets subsections that follow.

### *Application 1: The multiplicative component model*

The multiplicative expression of the saturated log-linear model, called the multiplicative component model, reproduces the elements of the flow table as follows:

$$n_{ij} = (T)(O_i)(D_j)(OD_{ij}). \quad (1)$$

Like all saturated models, it is, strictly speaking, not a model but a way of representing the data. $n_{ij}$ is the observed flow of migration from region $i$ to region $j$, and the effect parameters are $T$, $O_i$, $D_j$, $OD_{ij}$. Therefore, any $i$ to $j$ flow found in the interior 6 by 6 sub-matrices of Table 37.1 can be expressed by an equation of the same form as Equation 1 with the corresponding set of parameters. $T$ gives the overall effect, $O_i$ gives the effect of origin $i$, $D_j$ gives the effect of destination $j$, and $OD_{ij}$ gives the effect of the association between $O$i and $D_j$. Taken together, the parameters of the saturated model represent the spatial structure of migration (Rogers, Willekens, Little *et al.* 2002).

Two different sets of parameters that satisfy the multiplicative component model have been used in migration studies and both are presented here. Each one offers a different way of representing and interpreting the migration structure. The first is called geometric mean effect coding (Knoke and Burke 1980; Willekens 1983) and the second is called total sum reference coding (Raymer and Rogers 2007; Rogers, Little and Raymer 2010). A third multiplicative component model is derived in the subsection presenting the log-linear additive model.

### *Application 2: Geometric mean effect coding*

Geometric mean effect coding was the first decomposition of Equation 1 used for migration analysis. It was proposed by Birch (1963) and is formally equivalent to the gravity model of migration (Willekens 1983). Table 37.2 shows the multiplicative components resulting from geometric mean effect coding of the Netherlands data from Table 37.1. Note that the overall component ($T$) is set out in the grand total locations of the table, the origin components ($O_i$) are set out

in the row-total locations, the destination components ($D_j$) are set out in the column-total locations, and the origin-destination interaction components ($OD_{ij}$) are set out in the cells of the interior sub-matrices.

The overall effect, $T$, is described as the constant of proportionality or the *size* main effect (Willekens 1983). It is the geometric mean of all inter-regional flow values:

$$T = \left[ \prod_{ij} n_{ij} \right]^{\left(\frac{1}{m \times m}\right)},$$

where $m$ is the number of origin regions (rows) = the number of destination regions (columns). $T$ equals 17,168.003 for 1973 and 16,401.919 for 1976.

For a particular region $i$, the main effect of that region of origin is the ratio of the geometric mean of flows originating from $i$ divided by the overall geometric mean:

$$O_i = \frac{1}{T} \left[ \prod_j n_{ij} \right]^{\frac{1}{m}}.$$

The main effect, $O_i$, shows the relative importance of region $i$ as a source of migrations (Alonso 1986). For example, based on the 1973 data, the effect of originating in Category 4 is equal to:

$$O_4 = \frac{1}{17168.003} \left[ \begin{matrix} 23457 \times 14169 \times 4431 \times \\ 10209 \times 9386 \times 10973 \end{matrix} \right]^{\frac{1}{6}} = 0.627.$$

This is the smallest of the origin (row) effects, which suggests that Category 4 was the least important source of migrations in 1973.

Similarly, the destination main effect, $D_j$, gives the relative importance of region $j$ as an attractor of migrants. It is ratio

*TABLE 37.1 Migration between municipalities by degree of urbanization,* the Netherlands, 1973 and 1976*

**A. 1973 Migration table**

| | | | Destination | | | | |
|---|---|---|---|---|---|---|---|
| **Origin** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
| 1 | 50,498 | 23,829 | 8,566 | 21,846 | 16,264 | 18,856 | 139,859 |
| 2 | 25,005 | 27,536 | 6,953 | 14,326 | 16,212 | 18,282 | 108,314 |
| 3 | 15,675 | 10,710 | 13,874 | 6,266 | 9,819 | 19,701 | 76,045 |
| 4 | 23,457 | 14,169 | 4,431 | 10,209 | 9,386 | 10,973 | 72,625 |
| 5 | 29,548 | 25,267 | 11,802 | 13,160 | 15,979 | 20,406 | 116,162 |
| 6 | 46,815 | 39,123 | 42,399 | 25,012 | 26,830 | 23,304 | 203,483 |
| **TOTAL** | **190,998** | **140,634** | **88,025** | **90,819** | **94,490** | **111,522** | **716,488** |

**B. 1976 Migration table**

| | | | Destination | | | | |
|---|---|---|---|---|---|---|---|
| **Origin** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
| 1 | 14,473 | 14,327 | 6,077 | 11,689 | 10,618 | 9,897 | 67,081 |
| 2 | 14,833 | 36,258 | 13,289 | 17,391 | 20,899 | 21,869 | 124,539 |
| 3 | 8,330 | 17,764 | 25,113 | 10,489 | 18,171 | 29,220 | 109,087 |
| 4 | 11,315 | 16,498 | 8,935 | 10,537 | 10,762 | 12,519 | 70,566 |
| 5 | 11,875 | 24,370 | 19,151 | 12,312 | 16,724 | 22,591 | 107,023 |
| 6 | 16,582 | 32,336 | 52,415 | 22,264 | 28,182 | 27,810 | 179,589 |
| **TOTAL** | **77,408** | **141,553** | **124,980** | **84,682** | **105,356** | **123,906** | **657,885** |

*1: rural municipalities
2: industrial rural municipalities
3: specific resident municipalities of commuters
4: rural towns and small towns
5. medium-sized towns
6. large towns of more than 100,000 inhabitants

Source: Central Bureau of Statistics, The Hague

of the geometric mean of column $j$ to the total geometric mean and the formula is:

$$D_j = \frac{1}{T}\left[\prod_i n_{ij}\right]^{\frac{1}{m}}.$$

For example, for municipalities in Category 4, the destination effect in 1973 is equal to:

$$D_4 = \frac{1}{17168.00}\left[\begin{matrix}21846\times14326\times6266\times\\10209\times13160\times25012\end{matrix}\right]^{\frac{1}{6}} = 0.798.$$

All other row and column effects can be derived in the same way. Each is the geometric mean of the row (or column) elements divided by the overall geometric mean, and they are equivalent to the balancing factors in the gravity model (Willekens 1983).

The effects can be compared across regions and across time periods. For example, Category 6 was the most important source of migrations in 1973 (1.903 is greater than the other destination effects), and in 1976 (1.712 is greater than the other destination effects). Category 1 was less important as a destination in 1976 than in 1973 (0.768 is less than 1.711), and, in 1973, it was less important as a source of migrations than as a destination for migrations (1.180 is less than 1.711).

Panels A and B in Table 37.2 are sometimes called the spatial interaction matrices. The elements are the $OD_{ij}$ interaction effects in Equation 1 and each one is equal to the observed flow between $i$ and $j$ divided by the expected flow, which is the product of the other three parameters. The formula is:

$$OD_{ij} = \frac{n_{ij}}{(T)(O_i)(D_j)}.$$

Each $OD_{ij}$ expresses the departure of the observed flow, $n_{ij}$, from the expected flow based on the assumption of no association between the destination region $j$ and the origin region $i$, i.e., $(T)(O_i)(D_j)$. They have been interpreted as indicators of the accessibility, the ease of interaction, or the attractiveness between two regions (Rogers, Willekens, Little *et al.* 2002).

*TABLE 37.2 Multiplicative components using geometric mean effect coding*

| **A. 1973 Migration table** | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | **Destination** | | | |
| **Origin** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
| 1 | 1.457 | 0.940 | 0.656 | 1.352 | 0.933 | 0.882 | 1.180 |
| 2 | 0.885 | 1.332 | 0.653 | 1.087 | 1.140 | 1.048 | 0.962 |
| 3 | 0.771 | 0.720 | 1.811 | 0.661 | 0.959 | 1.570 | 0.692 |
| 4 | 1.275 | 1.052 | 0.639 | 1.190 | 1.014 | 0.966 | 0.627 |
| 5 | 0.943 | 1.102 | 1.000 | 0.901 | 1.013 | 1.055 | 1.067 |
| 6 | 0.838 | 0.957 | 2.015 | 0.960 | 0.954 | 0.676 | 1.903 |
| **TOTAL** | **1.711** | **1.252** | **0.644** | **0.798** | **0.861** | **1.056** | **17,168.003** |
| **B. 1976 Migration table** | | | | | | | |
| | | | | **Destination** | | | |
| **Origin** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
| 1 | 1.753 | 0.984 | 0.571 | 1.317 | 0.979 | 0.787 | 0.656 |
| 2 | 0.986 | 1.366 | 0.686 | 1.075 | 1.057 | 0.954 | 1.195 |
| 3 | 0.655 | 0.792 | 1.533 | 0.767 | 1.088 | 1.508 | 1.010 |
| 4 | 1.277 | 1.055 | 0.783 | 1.106 | 0.925 | 0.927 | 0.704 |
| 5 | 0.900 | 1.047 | 1.127 | 0.868 | 0.965 | 1.124 | 1.048 |
| 6 | 0.769 | 0.850 | 1.888 | 0.960 | 0.995 | 0.847 | 1.712 |
| **TOTAL** | **0.768** | **1.354** | **0.989** | **0.825** | **1.008** | **1.169** | **16,401.919** |

Values equal to 1.0 indicate independence, i.e., no association between the origin and the destination. As implied by Equation 1, if an $OD_{ij}$ parameter is equal to 1.0, $n_{ij}$ is determined by the values of $T$, $O_i$ and $D_j$ alone. A departure from 1.0 in either direction is an indication of an association between the destination and the origin. Values greater than 1.0 indicate higher than expected levels of accessibility/attractiveness and values less than 1.0 indicate less than expected accessibility/attractiveness.

Since the 1973 diagonal effects are generally greater than 1.0, it appears migrants were unexpectedly attracted to destinations in the same category of municipality. Category 6 was an exception. Migrants from large towns of more than 100,000 inhabitants (i.e., Category 6) were more attracted to commuter municipalities (i.e., Category 3) than to other large towns (2.015 is greater than 0.676).

Table 37.2 shows all the parameters necessary for reproducing the 1973 and 1976 flows. To verify that any flow in Table 37.1 can be reproduced by the multiplicative components, take, for example, the 1973 flow from Category 2 to Category 3:

$$n_{2,3} = 6953 = 17168.003 \times 0.962 \times 0.644 \times 0.653.$$

The parameter values, however, are not all independent of each other. In other words, some parameter values can be derived from the others. For one year of data, for all $i$ and $j$ combinations, there are 36 interaction effects, 6 origin main effects, 6 destination main effects, and one overall effect as reported in Table 37.2. However, the 49 parameters, reported for each year in Table 37.2, were derived from only 36 observed flows, implying that 13 parameters must be redundant. In other words, 13 of the 49 parameters can be calculated from the other 36, and the relationship between parameters is determined by the following constraints associated with geometric mean effect coding. The first set of constraints forces the products of the origin main effects (and destination effects) to be equal to 1. This is expressed as

$$\prod_i O_i = 1 \text{ and } \prod_j D_j = 1.$$

The second set of constraints is imposed on the interaction elements of each row and column, making the products of the interior elements in each row (and column) equal to 1. In other words, if five of the interaction effects associated with a particular origin (or destination) are given, the sixth interaction effect would be implied.

This is expressed as

$$\prod_i O_{ij} = 1 \text{ and } \prod_j D_{ij} = 1.$$

In general, if there are $m$ regions there are $m_2$ linearly independent parameters and $1 + m + m + (m \times m)$ multiplicative components. For all of the geometric mean effect coding computations, see Table 37.2 in the *Multiplicative Components* sheet of the accompanying workbook (see website).

***Application 3: Total sum reference coding***

Geometric mean effect coding, which uses the geometric mean as the reference value, was the earliest log-linear decomposition used to describe migration (Rogers, Willekens, Little *et al.* 2002; Willekens 1983). Recently, however, total sum reference coding has become more standard (Raymer and Rogers 2007; Rogers, Little and Raymer 2010). While both decompositions satisfy Equation 1, the effects under total sum reference coding are more transparent. For example, the main effect, $T$, is now the total number of migrants, denoted $n_{++}$. $O_i$ is now the proportion of all migrants leaving from region $i$ (i.e., $n_{i+}/n_{++}$), and $D_j$ is the proportion of all migrants moving to region $j$ (i.e., $n_{+j}/n_{++}$). The interaction component $OD_{ij}$ is now defined as $n_{ij}/[(T)(O_i)(D_j)]$ or the ratio of the observed number of migrants, $n_{ij}$, to the expected number, $(T)(O_i)(D_j)$. All effects taken together provide another way to represent the spatial structure of migration.

The multiplicative components derived from total sum reference coding are set out in Table 37.3. Consider, for example, the 8566 migrations from Category 1 to Category 3 in 1973 disaggregated into the four multiplicative components:

$$\begin{aligned}
n_{13} &= (T)(O_1)(D_3)(OD_{13}) \\
&= n_{++}\left(\frac{n_{1+}}{n_{++}}\right)\left(\frac{n_{+3}}{n_{++}}\right)\left[\frac{n_{13}}{\left(n_{++}\right)\left(\frac{n_{1+}}{n_{++}}\right)\left(\frac{n_{+3}}{n_{++}}\right)}\right] \\
&= (716488)\left(\frac{139859}{716488}\right)\left(\frac{88025}{716488}\right)\left(\frac{8566}{17183}\right) \\
&= 716488(0.102)(0.190)(0.477) \\
&= 8566.
\end{aligned}$$

The interpretations of these components are relatively straightforward. The overall component is the reported total number of migrations in 1973, i.e., 716,488. The origin component represents the share of all migrants from each region, i.e., 10 per cent of all migrations originated in the Category 1. The destination component represents the shares of all migrations to each region, i.e., 19 per cent of all migrations had Category 3 as the destination. Finally, the interaction component represents the ratio of observed migrants to expected migrants; thus there were roughly 48 observed migrations between region 1 and 3 for every 100 expected. The expected flow is based on the marginal total information, i.e., $(T)(O_1)(D_3)$.

Like geometric mean effect coding, the decomposition based on total sum reference coding gives more parameters than original data points. The constraints that define the relationships between parameters, and thus allow the redundant parameters to be derived, are as follows:

$$\sum_i O_i = 1;\ \sum_j D_j = 1;\ \frac{\sum_i O_i \sum_j OD_{ij}}{m} = 1;$$

and

$$\frac{\sum_j O_i \sum_i OD_{ij}}{m} = 1,$$

where $m$ is the number of regions (Raymer, Bonaguidi and Valentini 2006).

For all of the total sum reference coding computations, see Table 37.3 in the *Multiplicative components* sheet of the accompanying workbook (see website).

### *Comparing two multiplicative component models*

If the same decomposition scheme is applied to two sets of flow data from a given system of regions, all but the $T$ parameter are scale free. This means that taking the ratios of two sets of components provides a simple method for examining stability in migration structure without confounding the effects of growth or decline in overall levels of migration (Rogers, Willekens, Little *et al.* 2002). In Table 37.4, ratios of the 1976 to 1973 components are displayed. Several depart substantially from 1 indicating the migration structure changed in the three years between

*TABLE 37.3 Multiplicative components using total sum reference coding*

| **A. 1973 Migration table** | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | **Destination** | | | |
| **Origin** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
| 1 | 1.354 | 0.868 | 0.499 | 1.232 | 0.882 | 0.866 | 0.195 |
| 2 | 0.866 | 1.295 | 0.523 | 1.043 | 1.135 | 1.084 | 0.151 |
| 3 | 0.773 | 0.718 | 1.485 | 0.650 | 0.979 | 1.664 | 0.106 |
| 4 | 1.212 | 0.994 | 0.497 | 1.109 | 0.980 | 0.971 | 0.101 |
| 5 | 0.954 | 1.108 | 0.827 | 0.894 | 1.043 | 1.129 | 0.162 |
| 6 | 0.863 | 0.980 | 1.696 | 0.970 | 1.000 | 0.736 | 0.284 |
| **TOTAL** | **0.267** | **0.196** | **0.123** | **0.127** | **0.132** | **0.156** | **716,488** |
| **B. 1976 Migration table** | | | | | | | |
| | | | | **Destination** | | | |
| **Origin** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
| 1 | 1.834 | 0.993 | 0.477 | 1.354 | 0.988 | 0.783 | 0.102 |
| 2 | 1.012 | 1.353 | 0.562 | 1.085 | 1.048 | 0.932 | 0.189 |
| 3 | 0.649 | 0.757 | 1.212 | 0.747 | 1.040 | 1.422 | 0.166 |
| 4 | 1.363 | 1.087 | 0.667 | 1.160 | 0.952 | 0.942 | 0.107 |
| 5 | 0.943 | 1.058 | 0.942 | 0.894 | 0.976 | 1.121 | 0.163 |
| 6 | 0.785 | 0.837 | 1.536 | 0.963 | 0.980 | 0.822 | 0.273 |
| **TOTAL** | **0.118** | **0.215** | **0.190** | **0.129** | **0.160** | **0.188** | **657,885** |

1973 and 1976. For example, the ratio of the components for $OD_{11}$ is equal to 1.354, implying that migration within Category 1 was more attractive in 1976 than in 1973. In contrast, the ratio of the components for $OD_{33}$ is equal to 0.816, suggesting migration within Category 3 was less attractive in 1976 than in 1973.

***Application 4: The log-linear additive model***

Another form of the saturated log-linear model, which is an alternative to the multiplicative component model, is the linear additive model. Whether using the linear additive or the multiplicative form of the saturated log-linear model, the parameters represent the spatial structure of migration (Rogers, Willekens, Little *et al.* 2002) and each flow value can be fully reproduced by the parameters.

Because the multiplicative formation is formally equivalent to the gravity model (Willekens 1983), it is considered to be more appropriate than the linear additive model for representing spatial migration structures. On the other hand, the linear additive form is often found in statistics and when a standard statistical package (e.g., *SPSS*, *Stata*, *R*) is used to estimate a log-linear model, the parameters are always reported in the linear additive form. For that reason, the conventional calculations and interpretations of the parameters in the linear additive model are described in this sub-section.

The additive formulation is a linear function of logarithms and it makes evident why the model came to be called the log-linear model (Knoke and Burke 1980). It is mathematically equivalent to the multiplicative component model and it results from taking logarithms of both sides of Equation 1 as follows:

$$\ln(n_{ij}) = \ln(T) + \ln(O_i) + \ln(D_j) + \ln(OD_{ij})$$

which can be expressed more concisely as:

$$\ln(n_{ij}) = \lambda + \lambda_i^O + \lambda_j^D + \lambda_{ij}^{OD}. \quad (2)$$

The $\lambda$ values are simply the natural logarithms of the parameters appearing in Equation 1. The *O*, *D*, and *OD* superscripts are parameter descriptors (not exponents) and the subscripts *i* and *j* refer to the categories of the origin and destination variables, respectively.

Applying natural logarithmic transformations to the parameters in Table 37.2 and Table 37.3 would result in sets of corresponding linear additive parameters. However, just as there are at least two decompositions of the multiplicative component model, i.e., the geometric mean reference coding and the total sum effect coding, there are multiple strategies for arriving at sets of parameters that satisfy the linear additive model (Powers and Xie 2008), and the approaches taken by the standard statistical packages are not simply logarithmic transformations of the multiplicative components derived earlier.

Recall that a migration system with $m$ regions has $m \times m$ linearly independent parameters. The multiplicative component models described above give an interpretable value for $1 + m + m + (m \times m)$ parameters, though they are not linearly independent of each other. On the other hand, statistical routines in *SPSS*, *Stata*, and *R* calculate and report only linearly independent parameters, resulting in 1 value for $\lambda^T$, $m-1$ values for $\lambda_i^O$, $m-1$ values for $\lambda_j^D$, and $(m-1) \times (m-1)$ values for $\lambda_{ij}^{OD}$.

The particular set of parameter values that is calculated and reported depends on the contrast coding scheme used by the software. Contrast coding blocks out one region by fixing all linear additive parameters for that region equal

*TABLE 37.4 Ratios of 1976 to 1973 multiplicative components*

| | | | | Destination | | | |
|---|---|---|---|---|---|---|---|
| **Origin** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
| 1 | 1.354 | 1.144 | 0.957 | 1.099 | 1.121 | 0.904 | 0.522 |
| 2 | 1.169 | 1.045 | 1.075 | 1.040 | 0.923 | 0.860 | 1.252 |
| 3 | 0.839 | 1.055 | 0.816 | 1.149 | 1.062 | 0.854 | 1.562 |
| 4 | 1.125 | 1.093 | 1.342 | 1.046 | 0.972 | 0.970 | 1.058 |
| 5 | 0.988 | 0.955 | 1.139 | 1.000 | 0.936 | 0.993 | 1.003 |
| 6 | 0.909 | 0.854 | 0.906 | 0.993 | 0.980 | 1.117 | 0.961 |
| **TOTAL** | **0.441** | **1.096** | **1.546** | **1.015** | **1.214** | **1.210** | **0.918** |

to 0. *SPSS*, for example, fixes the parameters for the last region, i.e., the region assigned the highest numeric value, *m,* in this case: $\lambda_m^O = \lambda_m^D = \lambda_{mj}^{OD} = \lambda_{im}^{OD} = 0$. The parameters of the Netherlands data reported by *SPSS* are displayed in Table 37.5. The *SPSS* commands that generate these results for the 1973-migration table, along with the *SPSS* output, are presented in Appendix 1 (available on the *Tools for Demographic Estimation* website). Table 37.5 with the Excel formulae for calculation of the parameters is available in the *Contrast coding* sheet of the accompanying workbook (see website).

Notice the parameters for the last region are equal to 0, and, therefore, make no contribution to Equation 2. Interpretation of the parameters in Table 37.5 is somewhat complicated since they are in logarithmic units. Conversion back to the multiplicative components by exponentiation gives yet another set of multiplicative components that satisfy Equation 1. These are presented in Table 37.6, and they are the multiplicative components associated with "last region" contrast coding. Generally, these are not used to describe the spatial structure of migration, but they are useful in describing migration systems because the interaction parameters, $OD_{ij}$, are equivalent to odds ratios.

For example, the overall parameter from the 1973-migration data reported in Table 37.5, $\lambda^T$, gives the natural logarithm of the observed migrations for the reference region:

- $\ln(n_{66}) = 10.056$, and from Table 37.6, the companion parameter $T$ gives the $n_{66}$ migration flow:
  - $n_{66} = \exp(10.056) = 23304$.

Another illustration from the 1973-migration table in Table 37.5 shows how the origin main effects, $\lambda_i^O$, are added to the overall parameter to reproduce the migrations from Category 1 to the reference destination, Category 6, reported in Table 37.1. For example:

- $\ln(n_{16}) = 10.056 - 0.212 = 9.845$, and the corresponding multiplicative component, $O_1$ times $T$ from Table 37.6 gives:
  - $n_{16} = 27810 \times 0.356 = 18856$.

Using the same approach, the logarithms of all the migration flows can be reproduced by applying Equation 1 with the appropriate parameters from Table 37.6, or the observed

*TABLE 37.5 Additive linear parameters using "last region" contrast coding*

**A. 1973 Migration table**

| | | | Destination | | | | |
|---|---|---|---|---|---|---|---|
| **Origin** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
| 1 | 0.288 | –0.284 | –1.388 | 0.076 | –0.289 | 0.000 | –0.212 |
| 2 | –0.384 | –0.109 | –1.565 | –0.315 | –0.261 | 0.000 | –0.243 |
| 3 | –0.926 | –1.128 | –0.949 | –1.216 | –0.837 | 0.000 | –0.168 |
| 4 | 0.062 | –0.262 | –1.505 | –0.143 | –0.297 | 0.000 | –0.753 |
| 5 | –0.327 | –0.304 | –1.146 | –0.509 | –0.385 | 0.000 | –0.133 |
| 6 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **TOTAL** | **0.698** | **0.518** | **0.598** | **0.071** | **0.141** | **0.000** | **10.056** |

**B. 1976 Migration table**

| | | | Destination | | | | |
|---|---|---|---|---|---|---|---|
| **Origin** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
| 1 | 0.897 | 0.219 | –1.122 | 0.389 | 0.057 | 0.000 | –1.033 |
| 2 | 0.129 | 0.355 | –1.132 | –0.007 | –0.059 | 0.000 | –0.240 |
| 3 | –0.738 | –0.648 | –0.785 | –0.802 | –0.488 | 0.000 | 0.049 |
| 4 | 0.416 | 0.125 | –0.971 | 0.050 | –0.165 | 0.000 | –0.798 |
| 5 | –0.126 | –0.075 | –0.799 | –0.385 | –0.314 | 0.000 | –0.208 |
| 6 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **TOTAL** | **–0.517** | **0.151** | **0.634** | **–0.222** | **0.013** | **0.000** | **10.233** |

flows can be reproduced by applying Equation 2 using the parameters in Table 37.5.

The association parameters in the linear form, $\lambda_{ij}^{OD}$, are logged odds ratios (LORs), which are the logarithm of the ratio of two odds: 1) the odds of migration to region *j* rather than the reference region, conditional on originating in region *i*; and 2) the odds of migration to region *j* rather than the reference region, conditional on originating in the reference region. For example, from the 1973 sub-matrix in Table 37.5, $\lambda_{23}^{OD} = -1.565$, which is calculated as:

$$\lambda_{23}^{OD} = \ln\left[\frac{\frac{n_{23}}{n_{26}}}{\frac{n_{63}}{n_{66}}}\right] = \ln\left[\frac{\frac{6{,}953}{18{,}282}}{\frac{42{,}399}{23{,}304}}\right] = -1.565.$$

In words, the parameter is described as the logged ratio of the odds of migration to Category 3, rather than to Category 6, between a migrant originating in Category 2 and one originating in Category 6.

Odds ratios measure the relative likelihood of one outcome to another, and because they are more standard than LOR, it may be easier to exponentiate the LORs and interpret the association parameters, presented in Table 37.6, as odds ratios. For example, the model parameter $OD_{23}$, for the 1973 data, is calculated as:

$$OD_{23} = \exp(-1.565) = \left[\frac{\frac{n_{23}}{n_{26}}}{\frac{n_{63}}{n_{66}}}\right] = 0.209.$$

In words, the odds that a migrant from Category 2 will choose Category 3 over Category 6 is approximately 1/5th the odds that a migrant from Category 6 will choose Category 3 over Category 6. Odds-ratios are always positive and always depend on the choice of reference category. An odds ratio equal to 1 means a null relationship, i.e., statistical independence. Values higher than 1 mean a positive association and values less than 1 indicate a negative association.

*Stata* and *R* use a different contrast coding scheme to *SPSS*. Both of these statistical packages use the "first region" contrast coding as opposed to the "last region" contrast coding used by *SPSS*. In these two programs, the

*TABLE 37.6 Multiplicative components using "last region" contrast coding*

| **A. 1973 Migration table** | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | **Destination** | | | |
| **Origin** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
| 1 | 1.333 | 0.753 | 0.250 | 1.079 | 0.749 | 1.000 | 0.809 |
| 2 | 0.681 | 0.897 | 0.209 | 0.730 | 0.770 | 1.000 | 0.785 |
| 3 | 0.396 | 0.324 | 0.387 | 0.296 | 0.433 | 1.000 | 0.845 |
| 4 | 1.064 | 0.769 | 0.222 | 0.867 | 0.743 | 1.000 | 0.471 |
| 5 | 0.721 | 0.738 | 0.318 | 0.601 | 0.680 | 1.000 | 0.876 |
| 6 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| **TOTAL** | **2.009** | **1.679** | **1.819** | **1.073** | **1.151** | **1.000** | **23,304** |
| **B. 1976 Migration table** | | | | | | | |
| | | | | **Destination** | | | |
| **Origin** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
| 1 | 2.453 | 1.245 | 0.326 | 1.475 | 1.059 | 1.000 | 0.356 |
| 2 | 1.138 | 1.426 | 0.322 | 0.993 | 0.943 | 1.000 | 0.786 |
| 3 | 0.478 | 0.523 | 0.456 | 0.448 | 0.614 | 1.000 | 1.051 |
| 4 | 1.516 | 1.133 | 0.379 | 1.051 | 0.848 | 1.000 | 0.450 |
| 5 | 0.882 | 0.928 | 0.450 | 0.681 | 0.731 | 1.000 | 0.812 |
| 6 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| **TOTAL** | **0.596** | **1.163** | **1.885** | **0.801** | **1.013** | **1.000** | **27,810** |

parameters for the first region, i.e., the region assigned the lowest numeric value, are fixed to be equal to 0, i.e., $\lambda_1^O = \lambda_1^D = \lambda_{1j}^{OD} = \lambda_{i1}^{OD} = 0$. The *Stata* and *R* commands for generating the linear additive parameters, as well as the corresponding output, for the 1973 migration data can be downloaded from Appendix 1 to this chapter, (available on the *Tools for Demographic Estimation* website).

All forms of the saturated model and all statistical methods for estimating the interaction parameters are in agreement and provide substantively similar results. The formulae for the calculations of the parameters are available in the *Linear Additive Parameters* sheet of the accompanying workbook (see website). Furthermore, tests that each linear additive interaction parameter is equal to 0 are done automatically by *SPSS* and *Stata*. These results are available from Appendix 1 to this chapter (available on the *Tools for Demographic Estimation* website) and they show that each non-redundant interaction parameter is statistically significant. See Agresti and Finlay (2009) and Powers and Xie (2008) for descriptions of the standard errors of the estimates.

***Application 5: The independence model***

All the models presented to this point have been saturated, and, therefore, perfectly represent the observed flows. Generally, the substantively interesting parameters are the interaction parameters because they indicate associations between pairs of regions. The independence model, however, hypothesizes that the interaction parameters are uninteresting and unnecessary because all multiplicative interaction parameters, $OD_{ij}$, are equal to 1, or, equivalently, all linear additive interaction parameters, $\lambda_{ij}^{OD}$, are equal to 0. The independence model implies that the interaction terms should fall out of the model, reducing it to the most parsimonious form of a two-variable model, i.e., $n_{ij} = (T)(O_i)(D_j)$ or, equivalently, $\ln(n_{ij}) = \lambda + \lambda_i^O + \lambda_j^D$.

Visual inspection of the interaction parameters in the saturated log-linear model is one strategy for investigating the independence hypothesis. Another method is to calculate row or column conditional distributions. If the conditional distributions within rows (origins) are identical, there is independence between origins and destinations. In addition, since independence is a symmetric property, if the conditional distributions within rows (origins) are identical, the distributions within columns (destinations) also will be identical (Agresti and Finlay 2009; Powers and Xie 2008). In the *Independence* sheet of the accompanying workbook, the percentages of the Netherlands migrations within columns (destinations) are calculated. The column percentages are quite varied, suggesting, like the interaction parameters, that statistical independence is unfounded in this example.

The independence hypothesis implies that each particular inter-regional flow can be determined by the sizes of the marginal flows. Let $N_{ij}$ be the expected flow between regions $i$ and $j$ if the independence hypothesis is true. $N_{ij}$ is then equal to the total number of flows in the migration system, $n_{++}$, multiplied by the proportion of the all migrants leaving from region $i$, $n_{i+}/n_{++}$, times the proportion of all migrants moving to region $j$, $n_{+j}/n_{++}$, i.e., $N_{ij} = n_{++}(n_{i+}/n_{++})(n_{+j}/n_{++})$. If independence can be assumed, a good estimate of an inter-regional flow is $N_{ij}$, and the problem of estimating inter-regional migration flows is truly simplified.

The differences between the observed flows, $n_{ij}$, and the expected flows, $N_{ij}$, form the basis of the goodness-of-fit evaluation and the Pearson Chi-Squared Statistic, denoted $\chi^2$, which is widely used to summarize these discrepancies. It is calculated as: $\chi^2 = \sum \frac{(n_{ij} - N_{ij})^2}{N_{ij}}$, where the summation is taken over all internal cells in the migration matrix. When there is perfect agreement between the observed and the expected flows, over all cells, the $\chi^2$ equals 0 indicating the independence model fits the data perfectly. Larger differences between $n_{ij}$ and $N_{ij}$ produce larger $\chi^2$ values and increasingly stronger evidence that the independence model is inadequate. In general, smaller values indicate a good fit and larger values a poor fit.

If the independence hypothesis is true, the $\chi^2$ statistic is governed by the $\chi^2$ probability distribution with $(m-1)\times(m-1)$ degrees of freedom. This distribution provides the basis for testing the significance of the $\chi^2$ statistic (Agresti 2007; Agresti and Finlay 2009). If the $\chi^2$ statistic falls in the right-sided extremes of its distribution, it signifies a low probability, e.g., $p < 0.05$, that the independence hypothesis is true, and the model is rejected. The $\chi^2$ values associated with independence model applied to the Netherlands data in Table 37.1 are calculated and reported in the *Independence* sheet of the accompanying workbook. See Appendix 2 (available on the *Tools for Demographic Estimation* website) for the *SPSS*, *Stata* and *R* commands for testing the independence model with the 1973 example data.

The $\chi^2$ value associated with the 1973 example data is 47,623, and the degrees of freedom (*df*) are 25. The

associated *p*-value is less than 0.000, and the hypothesis of independence is rejected. (However, see the comments below about the limitations of this test when the sample size is large.) This is not surprising given the three multiplicative decompositions of the Netherlands data, presented in Table 37.2, Table 37.3 and Table 37.6. The evidence consistently shows strong associations between regions and many of the multiplicative association parameters are not close to 1. Furthermore, the standard errors reported in Appendix 1 to this chapter (available on the *Tools for Demographic Estimation* website) by *SPSS* and *Stata* indicate the linear additive interaction parameters are significantly different from 0.

One alternative to the $\chi^2$ statistic is called either the likelihood ratio statistic, the deviance, or the $G^2$ statistic. All are different names for the same test statistic, and which name is used is determined by the preferences of authors of text books and software packages. For simplicity, $G^2$ will be adopted here. It is similar to the $\chi^2$ in that values close to 0 indicate a well-fitting model and large values indicate a poor fit. If the hypothesized independence model holds, the $G^2$ statistic also has a $\chi^2$ distribution.

The $G^2$ statistic has general utility that goes well beyond the independence model in log-linear analysis. It is widely used for comparing a simpler model to a more complex model. The $G^2$ statistic is derived from the ratio of two likelihoods: 1) the likelihood that the constrained model (here the model of independence) fits the data; and 2) the likelihood that the unconstrained model (here the saturated model) fits the data. If the ratio is close to 1, the simpler, constrained, and more parsimonious model is preferred because it represents the data as well as the more complex model does.

The ratio of the two likelihoods does not have a $\chi^2$ distribution. However, when the ratio is transformed into natural logarithm units and multiplied by –2, it becomes $G^2$, which is a $\chi^2$ distributed variable with $(m-1)\times(m-1)$ degrees of freedom. If $L_c$ is the likelihood associated with the constrained (i.e., independence) model, and $L_u$ is the likelihood under the unconstrained (i.e., saturated) model, then $G^2$ is calculated as:

$$G^2 = -2\ln\left(\frac{L_c}{L_u}\right) = -2\ln L_c + 2\ln L_u \cdot$$

Because the saturated model fits the data perfectly (i.e., $L_u = 1$), $G^2 = -2\ln L_c$. The values, based on the example and the statistical software, are reported in Appendix 2 (see website). The value is reported to be 46,477.63 and it is called "Deviance" by *SPSS* and *Stata*. It is rounded and reported to be equal to 46,480 by *R*, where it is called "Residual Deviance." With 25 degrees of freedom the probability that the independence model holds is effectively 0.

The $\chi^2$ and the $G^2$ statistics are asymptotically equivalent (Powers and Xie 2008) and they form the bases of the Pearson Chi-square and the likelihood ratio tests, respectively. As with all inferential tests, effective use requires attention to underlying assumptions as well as limitations. Both tests rely on the assumption that each inter-regional flow count in the migration table follows an independent Poisson distribution (Powers and Xie 2008) and both tests have important limitations that are related to sample size. The $\chi^2$ statistic is inflated by large samples. Therefore, the Pearson Chi-square test is not appropriate when the sample size is large. The $G^2$ statistic and the likelihood ratio test is preferred in this situation (Powers and Xie 2008). The Pearson Chi-square test is preferred when the expected frequencies average between 1 and 10, but neither statistic works well if most of the expected frequencies are less than 5 (Agresti and Finlay 2009; Powers and Xie 2008).

Criticism has been made of the $G^2$ statistic as well when samples are large (Raftery 1986, 1995) and there is growing consensus that information measures should be considered along with traditional significance tests in assessing model fit. The Bayesian Information Criterion (*BIC*) is closely related to $G^2$, and it is calculated by *Stata* as: $BIC = G^2 - df\ln(m\times m)$, and by *SPSS* as:

$$BIC = -2\ln L_c + p\ln(m\times m),$$

where $p$ is the number of parameters estimated in the independence model, i.e., $2m-1$. A low value suggests choosing the independence model over the saturated model (Powers and Xie 2008).

Akaike's Information Criterion (AIC) is another alternative that takes on smaller values for better fitting models, since it judges how close the fitted values are to the expected values (Agresti 2007). In *SPSS* and *R*, it is calculated as:

$$AIC = -2(\ln L_c - p)$$

where $p$ is the number of parameters estimated in the independence model, i.e., $2m-1$. In *Stata*, it is calculated as:

$$AIC = \frac{-2(\ln L_c - p)}{m\times m}.$$

As shown in Appendix 2 (available on the *Tools for Demographic Estimation* website), *SPSS* and *Stata* report the *BIC* and *AIC*, and *R* reports only the rounded *AIC*. As previously stated, there are differences in the formulae used. The *BIC* reported by *SPSS* equals 46,934.237, and the *BIC* reported by *Stata* equals 46,388.04. *R* reports only the *AIC*, which is equal to 46,920, the rounded value reported by *SPSS*, 46,916.818. *Stata*'s *AIC* value is substantially smaller and is equal to 1,303.245. All reported *BIC* and *AIC* values are large and add to the growing evidence that discredits the independence model for this example.

**The quasi-independence model**

The independence model rarely provides an adequate fit to migration data. This is due, in part, to the overwhelming tendency to continue to reside in the same region. The quasi-independence model allows these "immobility" effects (Powers and Xie 2008) to be removed from the model, and this often results in improved predictions of inter-regional migration flows. The quasi-independence model has been applied effectively to migration data obtained from national censuses (Agresti 1990; Rogers, Little and Raymer 2010; Rogers, Willekens, Little *et al.* 2002), where persons who reported living in the same region at the time of the census as at the beginning of the reference period are represented in the diagonal elements of a migration table.

To illustrate, United States native-born migration data between 1985 and 1990 are reported in Panel A of Table 37.7. Clearly, the flows reported in the four diagonal elements of the interior sub-matrix are substantially larger than the off-diagonal elements, indicating that the propensity to maintain residence in the same region is much more typical than migration between regions.

The clustering along the diagonal cells contributes significantly to the poor fit of the independence model, and the dominating influence of the persons remaining in the region of origin have caused researchers to favour omitting them from the model. If migrants are defined as people changing their region of residence, this type of flow matrix is sometimes called a "migrants only" matrix. It is particularly useful for studying migration structure since it eliminates people who made no move or moved within the same region. Panel B of Table 37.7 displays the flow table with the diagonal elements set to 0, and the marginal totals adjusted accordingly.

The multiplicative components, using total sum reference coding, for the full migration table and the migrant-only table are reported in Table 37.8. The magnitude of the multiplicative component model parameters for the full data certainly departs from what is expected under the hypothesis of independence. They are substantially above 1.0 on the diagonal and the off-diagonal components are far below 1.0. In comparison, the multiplicative components for the migrants-only table are constrained to be equal to 0 in order to reproduce the structural zeros on the diagonal, and, as a result, the off-diagonal components are closer to 1.0.

*TABLE 37.7 United States native-born migration flows, 1985–1990*

| **A. Full migration table** | | | | | |
|---|---|---|---|---|---|
| | | | **Destination** | | |
| **Origin** | **Northeast** | **Midwest** | **South** | **West** | **Total** |
| Northeast | 40,262,319 | 336,091 | 1,645,843 | 479,819 | 42,724,072 |
| Midwest | 351,029 | 50,677,007 | 1,692,687 | 958,696 | 53,679,419 |
| South | 778,868 | 1,197,134 | 69,563,871 | 1,150,649 | 72,690,522 |
| West | 348,892 | 668,979 | 1,082,104 | 37,872,893 | 39,972,868 |
| **TOTAL** | **41,741,108** | **52,879,211** | **73,984,505** | **40,462,057** | **209,066,881** |
| **B. Migrants-only table** | | | | | |
| | | | **Destination** | | |
| **Origin** | **Northeast** | **Midwest** | **South** | **West** | **Total** |
| Northeast | 0 | 336,091 | 1,645,843 | 479,819 | 2,461,753 |
| Midwest | 351,029 | 0 | 1,692,687 | 958,696 | 3,002,412 |
| South | 778,868 | 1,197,134 | 0 | 1,150,649 | 3,126,651 |
| West | 348,892 | 668,979 | 1,082,104 | 0 | 2,099,975 |
| **TOTAL** | **1,478,789** | **2,202,204** | **4,420,634** | **2,589,164** | **10,690,791** |

The quasi-independence model requires that only migrations between different regions satisfy the independence assumption. This is estimated in two different but equivalent ways. The first method takes the full migration table data as in Panel A of Table 37.7, and fixes the weights on the interactive effects, $OD_{ij}$, to be zero when the regions of origin and destination are the same, i.e., $i=j$, insuring that $n_{ij}=0$. These are called structural zeros. When the origin and destination regions are different, i.e., $i \neq j$, the interaction effects are fixed at 1.0, which is the familiar independence model and gives the predicted off-diagonal flows under the quasi-independence hypothesis. Implementation of this method in *SPSS*, *Stata* and *R* is illustrated in Appendix 3 (available on the *Tools for Demographic Estimation* website).

The second method does not use the full migration data, but uses the migrants-only data as in Panel B of Table 37.7. It is best presented with the additive form: $\ln(n_{ij}) = \lambda + \lambda_i^O + \lambda_j^D + \delta_i I$, where $I$ is an indicator variable taking on values of 1 for the diagonal flows, i.e., when $i=j$, and values of 0 for the off-diagonal flows, i.e., when $i \neq j$ (Agresti 2002). Therefore, an extra parameter, $\delta_i$, is necessary to estimate each diagonal flow, and for the other inter-regional flows the $\delta_i I$ term falls out and the quasi-independence model reduces to the independence model. Consequently, just like the independence model, the off-diagonal interaction terms are constrained to be equal to 0 in the additive form of the model (and equal to 1 in the multiplicative form). Application of this method in *Stata* is illustrated in Appendix 3 (available on the *Tools for Demographic Estimation* website).

In the first method, the quasi-independence model fixes $m$ parameters, $OD_{ii}$, for $i=1$ to $m$, to be equal to 0. In the second method, $m$ additional parameters, $\delta_i$, are estimated, and when exponentiated will be very close to 0. Using either method, the quasi-independence model has $m$ more parameters than the full independence model and the degrees of freedom are reduced by $m$.

Appendix 3 (available on the *Tools for Demographic Estimation* website) illustrates how the quasi-independence model is estimated with statistical software packages *SPSS*, *Stata* and *R*, using the United States native-born migration flow data, 1985–1990. When the independence model is estimated with the full data, as expected, all goodness-of-fit indicators are extremely large: $\chi^2 = 544{,}479{,}395$ ($df=9$); $G^2 = 461{,}411{,}576$ ($df=9$); *Stata* values for *BIC* and *AIC* are 461,000,000 and 28,800,000, respectively. When the quasi-independence model is estimated, all values were reduced substantially: $\chi^2 = 327{,}233$ ($df=5$); $G^2 = 330{,}220$($df=5$); *Stata* values for *BIC* and *AIC* equal 330,207 and 27,535, respectively.

The inferential tests remain significant, and the quasi-independence model must be rejected as the true migration model. The independence and the quasi-independence models should not be compared, inferentially, with the likelihood ratio test because they are not nested models. However, the information measures may be compared directly. Both

**A. Full migration table**

| | Destination | | | | |
|---|---|---|---|---|---|
| **Origin** | **Northeast** | **Midwest** | **South** | **West** | **Total** |
| Northeast | 4.720 | 0.031 | 0.109 | 0.058 | 0.204 |
| Midwest | 0.033 | 3.733 | 0.089 | 0.092 | 0.257 |
| South | 0.054 | 0.065 | 2.704 | 0.082 | 0.348 |
| West | 0.044 | 0.066 | 0.076 | 4.896 | 0.191 |
| **TOTAL** | **0.200** | **0.253** | **0.354** | **0.194** | **209,066,881** |

**B. Migrants-only table**

| | Destination | | | | |
|---|---|---|---|---|---|
| **Origin** | **Northeast** | **Midwest** | **South** | **West** | **Total** |
| Northeast | 0.000 | 0.663 | 1.617 | 0.805 | 0.230 |
| Midwest | 0.845 | 0.000 | 1.363 | 1.318 | 0.281 |
| South | 1.801 | 1.859 | 0.000 | 1.520 | 0.292 |
| West | 1.201 | 1.547 | 1.246 | 0.000 | 0.196 |
| **TOTAL** | **0.138** | **0.206** | **0.413** | **0.242** | **10,690,791** |

*TABLE 37.8 Multiplicative components* of United States native-born migration flows, 1985–1990*

*Total sum reference coding

the *BIC* and *AIC* are reduced substantially, favouring the quasi-independence model over the independence model.

In addition, the predicted flows from the independence model can be contrasted with those from the quasi-independence model in Table 37.9. Visually comparing the predicted flows in Table 37.9 with the observed data in Table 37.7 reveals how much closer the quasi-independence model comes to representing the data. Two additional summary statistics are reported: $R^2$ and Mean Absolute Percent Error (*MAPE*). A comparison of the $R^2$ values shows the independence model explains 10 per cent of the variation in the observed data and the quasi-independence model explains 95 per cent. Furthermore, the average percent error for the quasi-independence model (*MAPE* = 28) is dramatically reduced in comparison to the independence model (*MAPE* = 2,492).

Since the fit of the quasi-independence model is not close enough to the observed data, it must be rejected as the "true" model. However, without observed migration data, the quasi-independence model may still offer a reasonable, but course, method for estimating inter-regional flows.

**Application 6: The method of offsets**

The validity of the independence and quasi-independence models can be evaluated with the inferential test statistics that accompany the log-linear model output, and, even when the models are not supported with significance tests, these models may be applied, in some contexts, to produce meaningful estimates of migration flows. The method of offsets assumes the auxiliary data have an implied structure of inter-regional associations that resembles the unknown migration structure. The method of offsets borrows the structure of the auxiliary data to derive the estimates of the missing migration flow data.

In past research, the auxiliary information, typically, has been a table of migration flows from another period in history (Rogers, Little and Raymer 2010; Rogers, Willekens, Little *et al.* 2002; Rogers, Willekens and Raymer 2003; Willekens 1983), but it could be from another age (Raymer and Rogers 2007), another sex or race group. It could be from another data source all together such as tax return data or motor vehicle registration data.

Given the auxiliary flow data, $n_{ij}^{*}$, the log-linear-with-offsets model is specified as: $\ln\left(\hat{n}_{ij}\right) = \lambda + \lambda_i^O + \lambda_j^D + \ln\left(n_{ij}^{*}\right)$. This model will estimate flows, $\hat{n}_{ij}$, that have a migration structure that comes as close as possible to that of the auxiliary flow data, and, at the same time, the estimated flows are adjusted to sum to the marginal totals pre-specified by the researcher. In this way, the method of offsets is similar to the independence and quasi-independence models in that it provides an expected distribution of the flows such that the marginal row and column totals are equal to the *a priori* estimates.

To illustrate the workings of the method of offsets, consider the Netherlands 1976 migration flow matrix in Table 37.1. Suppose we wish to keep the numerical values of

*TABLE 37.9 Predicted United States native-born migration flows, 1985–1990, under independence and quasi-independence*

| **A. Independence** | | | | |
|---|---|---|---|---|
| | | **Destination** | | |
| **Origin** | **1** | **2** | **3** | **4** |
| 1 | 8,530,046 | 10,806,184 | 15,119,178 | 8,268,664 |
| 2 | 10,717,328 | 13,577,116 | 18,996,052 | 10,388,923 |
| 3 | 14,512,977 | 18,385,588 | 25,723,693 | 14,068,264 |
| 4 | 7,980,756 | 10,110,323 | 14,145,583 | 7,736,206 |
| | | **$R^2$ = 0.104** | | ***MAPE* = 2492.322** |
| **B. Quasi-independence** | | | | |
| | | **Destination** | | |
| **Origin** | **1** | **2** | **3** | **4** |
| 1 | 0 | 535,839 | 1,349,561 | 576,353 |
| 2 | 442,768 | 0 | 1,793,640 | 766,005 |
| 3 | 720,681 | 1,159,163 | 0 | 1,246,806 |
| 4 | 315,340 | 507,201 | 1,277,434 | 0 |
| | | **$R^2$ = 0.945** | | ***MAPE* = 27.575** |

the row and column marginal totals, but, at the same time, wish to replace the migration interaction effects observed during that year by those observed during 1973, using the method of offsets. What would be the corresponding set of log-linear parameters? Table 37.10 sets out the predicted flow matrix obtained by the method of offsets in Panel A, and Panel B presents the associated multiplicative components derived using the total sum reference coding. Note that the $T$, $O_i$ and $D_j$ values of the predicted matrix, i.e., Panel B of Table 37.10, are identical to those reported for the observed 1976 flow matrix in Panel B of Table 37.3. However, the other terms (i.e., the interaction effects, $OD_{ij}$) reflect the influence of the migration structure of the observed 1973 data, Panel A of Table 37.3, as well as the row and column totals taken from the 1976 data. Therefore, the method of offsets applies the structure of the auxiliary data, the 1973 data in this case, to the interior flows, and at the same time, preserves the total number of flows observed in the 1976 data.

The predicted results in Panel A of Table 37.10 were taken from the output of the *SPSS*, *Stata*, and *R* commands for implementing the method of offsets found in Appendix 4 (available on the *Tools for Demographic Estimation* website). See the *Method of offsets* sheet in the accompanying Excel spreadsheet (see website) for other calculations.

Since the flows were observed directly in 1976, there are several ways to evaluate the suitability of the method of offsets for predicting the data. One simple method is to inspect visually the ratios of the association multiplicative components, as demonstrated in Table 37.4. Another method is to use the inferential tests and information measures reported by the log-linear procedures. These would be testing the hypothesis that the structure of the migration flows, i.e., the interaction parameters, did not change from 1973 to 1976. In the example reported in Table 37.10, the corresponding $G^2$ statistic is equal to 5,914 ($df = 25$), and the hypothesis that the auxiliary data represent the same migration process as the observed data must be rejected. The

***TABLE 37.10*** *Inter-regional migration flows in the Netherlands (1976), predicted with the method of offsets from the marginal totals (1976) and the migration flow table (1973)*

| **Panel A: Predicted using method of offsets** | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | **Destination** | | | | |
| **Origin** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
| 1 | 12,344 | 13,769 | 6,890 | 12,199 | 10,361 | 11,518 | 67,081 |
| 2 | 13,329 | 34,695 | 12,195 | 17,445 | 22,522 | 24,353 | 124,539 |
| 3 | 9,728 | 15,711 | 28,330 | 8,883 | 15,881 | 30,553 | 109,087 |
| 4 | 11,281 | 16,107 | 7,011 | 11,216 | 11,764 | 13,187 | 70,566 |
| 5 | 12,609 | 25,486 | 16,570 | 12,828 | 17,770 | 21,760 | 107,023 |
| 6 | 18,116 | 35,786 | 53,984 | 22,110 | 27,058 | 22,535 | 179,589 |
| **TOTAL** | **77,408** | **141,553** | **124,980** | **84,682** | **105,356** | **123,906** | **657,885** |
| | | | | | $R^2 = 0.966$ | | $MAPE = 8.364$ |
| **Panel B. Multiplicative components using total sum reference coding** | | | | | | | |
| | | | **Destination** | | | | |
| **Origin** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
| 1 | 1.564 | 0.954 | 0.541 | 1.413 | 0.964 | 0.912 | 0.102 |
| 2 | 0.910 | 1.295 | 0.515 | 1.088 | 1.129 | 1.038 | 0.189 |
| 3 | 0.758 | 0.669 | 1.367 | 0.633 | 0.909 | 1.487 | 0.166 |
| 4 | 1.359 | 1.061 | 0.523 | 1.235 | 1.041 | 0.992 | 0.107 |
| 5 | 1.001 | 1.107 | 0.815 | 0.931 | 1.037 | 1.080 | 0.163 |
| 6 | 0.857 | 0.926 | 1.582 | 0.956 | 0.941 | 0.666 | 0.273 |
| **TOTAL** | **0.118** | **0.215** | **0.190** | **0.129** | **0.160** | **0.188** | **657,885** |

final method, of those suggested here, relies on the standard $R^2$ and *MAPE* statistics to assess the fit between the observed and the predicted flows. These are reported in Panel A of Table 37.10 and are equal to 0.97 and 8.36, respectively. These statistics, as well as the ratios in Table 37.4, suggest this application of the method of offsets offers a set of estimates for the migration flows in 1976 that may be quite satisfactory.

The importance placed on the goodness-of-fit statistics depends on the quality of the observed flows used as inputs to the method of offsets. If the method is to be useful in a practical situation, it must be applicable when the inter-regional flows are not directly observed. In the absence of flow data, the method would still require pre-estimates of the marginal totals. Furthermore, if the method is implemented as illustrated in Appendix 4 (available on the *Tools for Demographic Estimation* website), initial estimates of the inter-regional flows are required. Therefore, the pre-estimates of the row and column totals would need to be distributed across the internal cells of the flow matrix so they add up to the respective marginal totals. Table 37.11, Panel A, presents a typical scenario, albeit continuing to use the marginal totals from the Netherlands 1976 data, which were observed. A simple solution is to distribute the flows according to the independence model, i.e., $\hat{n}_{ij} = (T)(O_i)(D_j)$, which results in the initial estimates of the flows displayed in Panel B of Table 37.11.

As long as the initial inter-regional flows add up to the marginal totals, the predicted flows are not affected by the method used to distribute the flows within the cells. This is true because the flows will be predicted, ultimately, from the auxiliary data through the method of offsets, using the iterative proportional fitting algorithm (Agresti 1990; Deming and Stephan 1940). In other words, the initial estimates of the 1976 Netherland flows, used as input to the offsets log-linear model, could be the internal cells of Table 37.1, Panel B, or those in Table 37.11, Panel B. Either set of initial estimates would yield the predicted flows that are reported in Table 37.10, Panel A.

On the other hand, it is important to note that the associated inferential test statistics and the information measures that accompany the method of offsets must be

*TABLE 37.11 The inputs to the method of offsets in the absence of observed flows*

| **Panel A. Pre-estimation marginal totals from the Netherlands, 1976** | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | **Destination** | | | | |
| **Origin** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
| 1 | | | | | | | 67,081 |
| 2 | | | | | | | 124,539 |
| 3 | | | | | | | 109,087 |
| 4 | | | | | | | 70,566 |
| 5 | | | | | | | 107,023 |
| 6 | | | | | | | 179,589 |
| **TOTAL** | **77,408** | **141,553** | **124,980** | **84,682** | **105,356** | **123,906** | **657,885** |
| **Panel B. Independence model distribution scheme for initial flow estimates** | | | | | | | |
| | | | **Destination** | | | | |
| **Origin** | **1** | **2** | **3** | **4** | **5** | **6** | **Total** |
| 1 | 7,893 | 14,433 | 12,744 | 8,635 | 10,743 | 12,634 | 67,081 |
| 2 | 14,654 | 26,796 | 23,659 | 16,030 | 19,944 | 23,456 | 124,539 |
| 3 | 12,835 | 23,472 | 20,724 | 14,042 | 17,470 | 20,545 | 109,087 |
| 4 | 8,303 | 15,183 | 13,406 | 9,083 | 11,301 | 13,290 | 70,566 |
| 5 | 12,593 | 23,027 | 20,331 | 13,776 | 17,139 | 20,157 | 107,023 |
| 6 | 21,131 | 38,641 | 34,117 | 23,116 | 28,760 | 33,824 | 179,589 |
| **TOTAL** | **77,408** | **141,553** | **124,980** | **84,682** | **105,356** | **123,906** | **657,885** |

interpreted with respect to the initial flow estimates. For example, if the initial flows were taken from Panel B of Table 37.11, the associated $\chi^2$ and $G^2$ test statistics would be testing the hypothesis that the predicted data are distributed in a manner that is consistent with the independence model.

It is a simple matter to modify the method of offsets to apply it to the problem of predicting a table of "migrants only." The *SPSS*, *Stata* and *R* commands require minor modifications that are specified in comments in Appendix 4 (available on the *Tools for Demographic Estimation* website). A worked example is included in the *Method of offsets, migrants only* sheet of the accompanying workbook (see website). It uses the observed US flows, 1985–1990, to retrospectively estimate the 1975-80 migrant flows reported by Rogers, Willekens, Little *et al.* (2002).

## REFERENCES


Agresti A. 1990. *Categorical Data Analysis.* New York: Wiley.

Agresti A. 2002. *Categorical Data Analysis.* New York: Wiley-Interscience.

Agresti A. 2007. *An Introduction to Categorical Data Analysis.* Hoboken, NJ: Wiley-Interscience.

Agresti A and B Finlay. 2009. *Statistical Methods for the Social Sciences.* Upper Saddle River, NJ: Pearson Prentice Hall.

Alonso W. 1986. *Systemic and log-linear models: From here to there, then to now, and this to that*. Discussion paper 86–10. Cambridge, MA: Harvard University, Center for Population Studies.

Birch MW. 1963. "Maximum likelihood in three-way contingency tables", *Journal of the Royal Statistical Society Series B-Statistical Methodology* **25**(1):220–233.

Deming WE and FF Stephan. 1940. "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known", *Annals of Mathematical Statistics* **11**(4):427–444. doi: 10.1214/aoms/1177731829

Knoke D and PJ Burke. 1980. *Log-linear Models.* Beverly Hills, CA: Sage Publications.

Mueser P. 1989. "The spatial structure of migration: An analysis of flows between states in the USA over three decades", *Regional Studies* **23**(3):185–200. doi: 10.1080/00343408912331345412

Nair PS. 1985. "Estimation of period-specific gross migration flows from limited data: Bi-proportional adjustment approach", *Demography* **22**(1):133–142. doi: 10.2307/2060992

Powers DA and Y Xie. 2008. *Statistical Methods for Categorical Data Analysis.* Bingley, UK: Emerald.

Raftery AE. 1986. "Choosing models for cross-classifications", *American Sociological Review* **51**(1):145–146. doi: 10.2307/2095483

Raftery AE. 1995. "Bayesian model selection in social research", *Sociological Methodology* **25**(1):111–163. doi: 10.2307/271063

Raymer J. 2007. "The estimation of international migration flows: A general technique focused on the origin-destination association structure", *Environment and Planning A* **39**(4):985–995. doi: 10.1068/a38264

Raymer J, A Bonaguidi and A Valentini. 2006. "Describing and projecting the age and spatial structures of interregional migration in Italy", *Population, Space and Place* **12**(5):371–388. doi: 10.1002/psp.414

Raymer J and A Rogers. 2007. "Using age and spacial flow structures in the indirect estimation of migration streams", *Demography* **44**(2):199–223. doi: 10.1353/dem.2007.0016

Rees P and FJ Willekens. 1986. "Data and accounts," in Rogers, A and FJ Willekens (eds). *Migration and Settlement: A Multi-regional Comparative Study*. Dordrecht: D. Reidel, pp. 19–58.

Rogers A, JS Little and J Raymer. 2010. *The Indirect Estimation of Migration: Methods for Dealing with Irregular, Inadequate, and Missing Data.* Dordrecht: Springer.

Rogers A, F Willekens, JS Little and J Raymer. 2002. "Describing migration spatial stucture", *Papers in Regional Science* **81**(1):29–48.

Rogers A, FJ Willekens and J Raymer. 2003. "Imposing age and spatial structures on inadequate migration-flow datasets", *The Professional Geographer* **55**(1):56–69.

Snickars F and JW Weibull. 1977. "A minimum information principle: Theory and practice", *Regional Science and Urban Economics* **7**(1–2):137–168. doi: 10.1016/0166-0462(77)90021-7

Willekens F. 1983. "Log-linear modeling of spatial interaction", *Papers of the Regional Science Association* **52**:187–205. doi: 10.1007/BF01944102